IEEE Ukraine Section

IEEE Ukraine Section NANO Council Chapter

IEEE Ukraine Section SP/AES Societies Joint Chapter

IEEE Ukraine Section (West) AP/ED/MTT/EP/SSC Societies Joint Chapter

IEEE Ukraine Section (Kharkiv) SP/AP/C/EMC/COM Societies Joint Chapter

Ivan Franko National University of Lviv

2023 IEEE 13th International Conference on ELECTRONICS AND INFORMATION TECHNOLOGIES (ELIT)

PROCEEDINGS

September 26 – 28, 2023

Lviv, Ukraine

ORGANIZERS:

- IEEE Ukraine Section (Ukraine)
- IEEE Ukraine Section NANO Council Chapter (Ukraine)
- IEEE Ukraine Section SP/AES Societies Joint Chapter (Ukraine)
- IEEE Ukraine Section (West) AP/ED/MTT/EP/SSC Societies Joint Chapter (Ukraine)
- IEEE Ukraine Section (Kharkiv) SP/AP/C/EMC/COM Societies Joint Chapter (Ukraine)
- Ivan Franko National University of Lviv (Ukraine)

PARTNER:

- G. V. Karpenko Physico-Mechanical Institute of the NAS of Ukraine (Ukraine)

2023 IEEE 13th International Conference on
Electronics and Information Technologies (ELIT)

Part Number CFP23LIT-ART

ISBN 979-8-3503-8309-6

ORGANIZING COMMITTEE

Velhosh S., Dr., Lviv, Ukraine (Chairman)

Karbovnyk I., Prof, Lviv, Ukraine (TPC-Chairman)

Kushnir O.O., Dr., Lviv, Ukraine (Publishing Chairman)

Katerynchuk I., Dr., Lviv, Ukraine (Financial Chairman)

Azarova I., Lviv, Ukraine

Bovgyra O., Dr., Lviv, Ukraine

Boyko Ya., Dr., Lviv, Ukraine

Dubyk V., Dr., Lviv, Ukraine

Dzendzelyuk O., Lviv, Ukraine

Kofluk I., Lviv, Ukraine

Koplak O., Lviv, Ukraine

Levush P., Lviv, Ukraine

Medvid I., Dr., Lviv, Ukraine

Shevchuk V., Lviv, Ukraine

Shmygelsky Ya., Lviv, Ukraine

Slobodzian D., Dr., Lviv, Ukraine

Stolyarchuk O., Lviv, Ukraine

Yaroshko S., Dr., Lviv, Ukraine

PROGRAMME COMMITTEE

Bolesta I., Prof., Lviv, Ukraine

Bordun O., Prof., Lviv, Ukraine

Buhrii O., Prof., Lviv, Ukraine

Chornodolskyy Ya., Dr., Lviv, Ukraine

Dyiak I., Prof., Lviv, Ukraine

Fitio V., Prof., Lviv, Ukraine

Furgala Yu., Dr., Lviv, Ukraine

Golovchak R., Prof., Clarksville, USA

Klym H., Prof., Lviv, Ukraine

Kukharskyy V., Dr., Lviv, Ukraine

Kushnir O.S., Prof., Lviv, Ukraine

Liashkevych V., Dr., Lviv, Ukraine

Melnyk B., Dr., Lviv, Ukraine

Monastyrskyi L., Prof., Lviv, Ukraine

Muravsky L., Prof., Lviv, Ukraine

Mykhaylyk V., Prof., Didcot, UK

Myshchyshyn O., Dr., Lviv, Ukraine

Nazarkevych M., Prof., Lviv, Ukraine

Ogirko I., Prof., Lviv, Ukraine

Olenych I., Prof., Lviv, Ukraine

Pavlyk B., Prof., Lviv, Ukraine

Pavlyshenko B., Prof., Lviv, Ukraine

Peleshko D., Prof., Lviv, Ukraine

Popov A., Dr., Riga, Latvia

Rendziniak S., Prof., Lviv, Ukraine

Rusyn B., Prof., Lviv, Ukraine

Shevchuk I., Prof., Lviv, Ukraine

Shuwar R., Dr., Lviv, Ukraine

Stasyshyn A., Prof., Lviv, Ukraine

Tsmots I., Prof., Lviv, Ukraine

Venherskyi P., Prof., Lviv, Ukraine

Vynokurova O., Prof., Lviv, Ukraine

Yarema O., Dr., Lviv, Ukraine

Yuzevych V., Prof., Lviv, Ukraine

TABLE OF CONTENTS

Attribute Selection, Outliers Impact Study and Visualization within Breast Cancer Detection

Gennady Chuiko
Computer Engineering Department
Petro Mohyla Black Sea National University
Mykolayiv, Ukraine
genchuiko@gmail.com

Olga Dvornik
Computer Engineering Department
Petro Mohyla Black Sea National University
Mykolayiv, Ukraine
https://orcid.org/0000-0002-4545-1599

Yeugen Darnapuk
IEEE member # 95367017, Ukraine Section
Computer Engineering Department
Petro Mohyla Black Sea National University
Mykolayiv, Ukraine,
yevhen.darnapuk@chmnu.edu.ua

Denis Honcharov
Information computer center
Petro Mohyla Black Sea National University
Mykolayiv, Ukraine
nemuritor@chmnu.edu.ua

Yaroslav Krainyk
Computer Engineering Department
Petro Mohyla Black Sea National University
Mykolayiv, Ukraine
https://orcid.org/0000-0002-7924-3878

Olga Yaremchuk
Medical Institute
Petro Mohyla Black Sea National University
Mykolayiv, Ukraine
olga.yaremchuk.77@ukr.net

***Abstract* — Classification of mammography data into two types of breast tumors, benign or malignant, is an effective screening tool and the primary way of diagnosis and decision-making. This report aims to opt for the most relevant attributes of the well-known Wisconsin Breast Cancer Diagnostic Data Set to reduce its size at first. The reduction was performed initially by ranking attributes and finally by "decision tree" analysis. The clipped data set had only six attributes, against 31 in the initial one. The five most relevant attributes were the following: "perimeter_worst," "area_worst," "concave points_worst," "texture_mean," and "concave points_mean." If possible, It should be done without losing classification potency. Over and above, our extra goal was to find classifiers that provide acceptable performance while allowing visualization of the results in a way accessible to clinicians. Here, we mean various visualization tools in the Machine learning framework: "decision trees," association rules, attribute ranking, and so forth, to improve breast cancer diagnosis.**

***Keywords* — Machine Learning, Breast Cancer Diagnostics, Attribute Selection, Decision-Trees, outliers**

I. Introduction

Breast cancer (BC) is the most prevalent cancer in Ukrainian women [1]. Women are the principal holders of the national gene pool. Therefore, BC early diagnosis with the help of mammography and artificial intelligence [2] calls for the stable interest of relevant experts and clinicians. In the last three decades, the gold standard in this area set the well-known Wisconsin Breast Cancer Diagnostic Dataset (WBCD) [3]. WBCD, in particular, allows for the forecast of BC by Machine Learning methods [4, 5, 6, and 7].

However, almost all papers about Machine Learning (ML) for diagnostics of BC, among which the cited above studies are only the "top of the iceberg," have specific common faults. The key is that they are over-technical and hard to interpret for clinicians unfamiliar with ML. Clinicians prefer simple and visual methods, like the "Decision tree" or associative rules, while many authors from ML society focus on high-performance ML algorithms [7], ignoring the visuality.

WBCD is detailed and reliable but a pretty bulky database: it comprises 569 instances with 31 attributes (features), including diagnostics, for each of them. Sure, these functions must catch out different worth (the relevance). So, specific attribute selection, as done in [5], and their ranking using various methods also seem reasonable. ML experts point out that such a reduced data set often performs better than the whole [5, 8, 9]. In addition, fewer attributes mean less laboring and faster or more timely analysis for mammography clinicians.

Thus, the goals of this study can be put in the form of two following points:

- Reduction of WBCD by dully selecting attributes and via other ML means.
- Selection of ML classifiers that provide acceptable performance while allowing visualization or at least simple interpretation, accessible for clinicians.

II. Methodology

A. Data set description

WBCD comprises ten descriptive geometrical parameters of tumors' nuclei extracted from mammography [3,4]. Those are:

- radius (mean of distances from the center to points on the perimeter),
- texture (standard deviation of gray-scale values),
- perimeter,
- area,
- smoothness (local variation in radius lengths),
- compactness (quadrate of perimeter/area),
- concavity (severity of concave portions of the contour),
- concave points (number of concave portions of the contour),
- symmetry,
- fractal dimension (via "coastline approximation").

Each image (mammogram) gives these numeric features from cell nuclei. Smaller sizes of tumors mean the earlier stage of BC. Thus, the WBCD data define the tumor's nature (benign or malignant) and the stage of breast cancer if the illness is set. Fig.1 shows the survival percentage and rough treatment costs depending on the BC stage. Both are plain tied with the tumor sizes.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

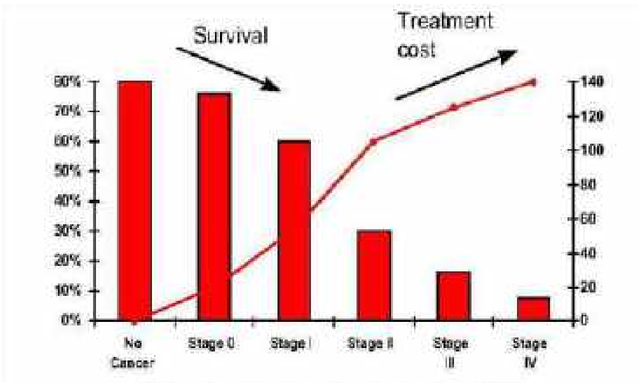


Fig. 1. Survival and BC treatment costs (the source: https://www.slideserve.com/lindsey/breast-cancer)

The list above holds ten geometric parameters. Why, then, does the WBCD contain up to 31 attributes? The point is that each of the ten above values is a three-component vector. The mean, standard error, and "worst" value (those are outliers, to be sure) of these features should be obtained for each one. There are 30 features (31 with diagnosis) as a result. Let us again stress that in this context, the "worst" means outliers (largest or smallest values). Thus, the outliers are one-third of the attributes within the initial WBCD.

B. *Data mining software*

Waikato Environment for Knowledge Analysis (Weka) is a Java-based software developed at the University of Waikato, New Zealand, and is free licensed under the GNU General Public License [8, 9]. The purpose is the data, particularly the big one, mining. The last versions of Weka are a modern collection of various algorithms and means of ML with powerful visuality support.

Namely, high-grade visualization provides the eye-catching usability of Weka for bio-medical data mining for clinicians. The well-developed Graphic User Interface (GUI) within Weka permits effective operation with data for one who is not an expert in Java or ML.

Weka works with a unique dataset file format (ARFF - Attribute Relations File Format) but also accepts ordinary CSF (comma-separated files) [9], in which the original WBCD was presented at first. We had the last version of Weka (3-9-6) in use. Note that this software is still not widespread enough among Ukrainian researchers. This report is one of the attempts to change this awkward situation.

III. RESULTS

As mentioned above, the original WBCD contains 569 copies. Each has 30 numerical attributes and one nominal (class), also called the diagnosis attribute (benign or malignant tumor: 357 and 212 cases, respectively). One can see that the two classes are pretty well-balanced in the data set.

Considering the given diagnoses, visual analysis of attributes histograms confirms the hypothesis of a higher probability of malignant cases with larger tumor sizes on average. The possibility of such visual analysis of attributes histograms is one of the advantages of Weka.

A. *Attributes selection and ranking*

The selection of attributes algorithms of Weka allows one to select the most relevant subset for both classes. That is a smaller but quite weighty subset among all features. Weka offers several algorithms for feature selection [9]. The Correlation-based Feature Selection (CFS) subset evaluator, which included locally predictive attributes, was used here. This evaluator accompanied the bi-directional "Best first" search method. The search closes itself after five node expansions. Total number of subsets evaluated: 420. The merit of the best subset found was 0.667.

The selected subset comprises as a result 11 attributes (12 with the diagnosis):

- two attributes of the "Standard_Error" type: "area_se" and "symmetry_se"(IDs 4 and 5);
- three attributes of the "Mean" type: “texture_mean,” “concavity_mean,” and “concave points_mean.” (IDs 1, 2, and 3);
- six attributes of the "Worst" type that is outliers: "radius_worst,"perimeter_worst," "area_worst," "smoothness_worst," "concavity_worst," and "concave points_worst." (IDs 6, 7, 8, 9,1 0, and 11)

Note that outliers now occupy more than half of the picked-up subset. The subset's attributes were ranked via the "Information gain attribute evaluator" in the range (of 0.023 up to 0.685). Four of the six characteristics of the "Worst" type have the four highest ranks (Fig.2)

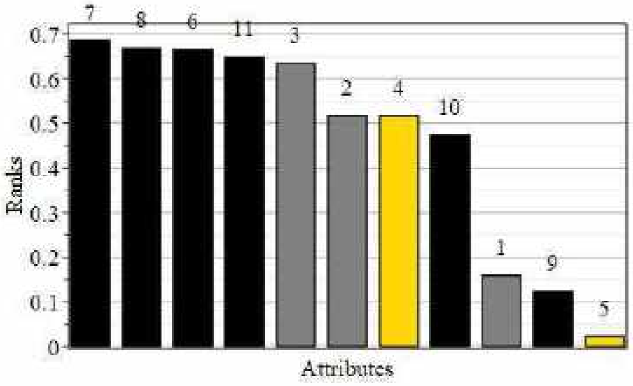


Fig. 2. Attributes ranking: the number (IDs) of attributes is shown over each column; the black color matches the "Worst"-type attributes, the grey - is the "Mean"-type, and the gold - is the "Standard_Error"-type.)

Fig.2 can hint at three groups of attributes: high-ranked (five first), mid-ranked (three), and low-ranked (last three). For instance, the authors [5] worked with a reduced data set with five attributes. Unfortunately, it is hard to identify which attributes were included in this subset from their text.

Farther, we will operate with all eleven attributes. Nonetheless, we will return to the problem of further data subset reduction and define our five-attribute subset below. We will make it with explicit attribute declarations, contrasting [5].

B. *Weka experiment: optimal classifier and its performance for the entire data set and reduced one*

Weka allows, after some customization with Package Manager, to apply the "Auto-Weka," a tool that performs combined algorithm selection and hyperparameter optimization. Auto-WEKA recommends a classifier for a specific dataset, which will likely serve best. Such an optimal classifier turned out to be the Logistic Models Tree (LMT) one in our research. Table 1 compares some performance parameters of this classifier with the best classifier of reference [7] (Support Vector Machine, SVM-type classifier).

Here, MCC means Mattew correlation coefficient. One can conclude that the performance is almost equally excellent for both classifiers.

Table 2 compares the performance parameters of the optimal LMT classifier concerning complete and reduced datasets. There are also given evaluations concerning both classes (diagnoses).

TABLE I. AVERAGE WEIGHTED PERFORMANCE ESTIMATIONS FOR TWO DIFFERENT CLASSIFIERS OF THE FULL DATASET

Classifier	Precision	Recall	Kappa statistics	MCC	ROC area
SVM	0.98	0.98	0.95	0.96	0.97
LMT	0.98	0.98	0.95	0.95	1.00

One can conclude that the dataset may be almost triply reduced due to the duly attribute selection. Meanwhile, the performance parameters of the reduced dataset are at least not worse than the ones for the complete dataset. Besides, the LMT classifier, recommended by Auto-Weka, ensures excellent performance, as shown in the tables above.

TABLE II. PERFORMANCE EVALUATIONS FOR COMPLETE AND REDUCED DATASETS

Data set	Class	Precision	Recall	Kappa	MCC	ROC area
Full	**M**	0.971	0.958	0.943	0.944	0.994
	B	0.975	0.983	0.943	0.944	0.994
Red.	**M**	0.985	0.963	0.951	0.951	0.993
	B	0.983	0.992	0.951	0.951	0.993

It is worth remarking that further reduction of the already reduced dataset, e.g., neglecting the three last low-ranked attributes, drops the classifying performance. The precisions decrease to 0,960 and 0.943, recall up to 0.901 and 0.978 for malignant and benign diagnoses, respectively, the MCC up to 0.891, and the ROC area up to 0.988. These are still excellent results but are a touch worse than those from Table II with eleven attributes dataset.

C. Clusters-to-classes evaluations: complete and reduced datasets

One can find a sample of the recent results of the WBCD clustering in the source [4]. Clustering was performed by two methods (K-means and Hierarchical method) concerning the full dataset only. The results of the work of the K-means algorithm were reckoned as more realistic. Our goal is slightly different: to evaluate cluster-to-classes' congruence for full and reduced datasets with the simple K-means algorithms. This algorithm allows two distances between instances: Euclidean [4] or Manhattan, which we had chosen.

TABLE III. COMPLEXITY (CONFUSION) MATRICES OF THE K-MEANS TECHNIQUE FOR FULL DATASETS

Results of [4]			Our results		
	M	**B**		**M**	**B**
Malignant	179	31	**Malignant**	184	28
Benign	10	349	**Benign**	9	348

Table III shows the complexity (alias confusion) matrices for cluster-to-classes congruence regarding our results and [4] ones. One can see that the percentage of incorrectly clustered instances is even slightly lower in our results: 6.5% against 7.2 % in [4]. So, Manhattan distance seems preferable for the K-means technique concerning WBCD. Still, the matching of results is good enough.

TABLE IV. COMPLEXITY (CONFUSION) MATRIX OF THE K-MEANS TECHNIQUE FOR THE REDUCED DATASET.

	Malignant	Benign
Malignant	189	23
Benign	3	354

Table IV gives the complexity matrix for clustering our reduced dataset. Unexpectedly, the result is even better than in the previous table. In particular, the percentage of incorrectly clustered instances is less than 4.6%. It means clusters and diagnosis (classes) match each other more than in 95% of cases, which is a reasonably okay result. Meanwhile, further reducing the subset, if one can neglect three low-ranked attributes, raises the percentage of incorrectly clustered instances to 5.3 % vs. the previous 4.6%.

Comparable but slightly worse results give algorithms EM (Expectation Maximization) and Density-based one: 5.3% and 6.3 % of incorrectly clustered instances, respectively. One shall mean the reduced dataset with eleven attributes within this paragraph.

D. Association Rules Mining

Association Rules mining helps us find patterns in the dataset [11]. It is the finding of attributes that occur together or are correlated if some simplifying is applied. Support and confidence are the leading measures for association rules mining.

Support measures the number of cases, or the probability, that two items occur together in a single transaction (e.g., bread and butter). Confidence (accuracy) is a measure that states the likelihood that two things arise one after the other but not together (e.g., laptop and antivirus software.) Support and confidence values limit the possible transactions and determine those most frequently occurring patterns [11]. One can find other, more special, measures of the association rules quality in [12].

TABLE V. EXAMPLES OF A FEW ASSOCIATION RULES FOR DIAGNOSES

Benign		
Conditions	**Support (probab.)**	**Confidence**
texture_mean < 19.5667, concave points_worst < 0.097	192 (0.538)	0.995
texture_mean < 1 9.5667, radius_worst < 17.3, concave points_worst < 0.097	189 (0.520)	0.993
Malignant		

perimeter_worst=(117.34-184.27) ,
smoothness_worst=(0.1216-0.1721)
	117
	(0.552)
		0.995

concave points_mean=(0.0671-0.1341), perimeter_worst=(117.34-184.27)	103 (0.485)	0.995

The reduced data set with eleven attributes was in use. Numeric attributes were discretized by three bins, which held low, mid, and high values. Weka has several algorithms for association rules mining. "Predictive Apriori" was used, which searches with an increasing support threshold for the best set of rules [13]. This algorithm has a handy option enabling class (diagnosis) association rules to be mined instead of standard association rules. Table V shows some of the most supported diagnoses.

Note that association rules for a benign diagnosis are based on conditions "less than" type that is, low values of attributes. At the same time, the conditions for malignant diagnosis use at least mid-attribute values.

E. Decision Tree and further reduction of the data set

Alas, the best classifier for the reduced data set (SGD, which exploits a Support Vector Machine type classification) does not, like the LMT, the "Decision Tree." Therefore, we have exploited a slightly tuned J48 classifier, which ensures only a touch worse performance than both classifiers mentioned above. For comparison, an averaged weighted precision of tuned J48 is 0.967, MCC is equal to 0.929, and ROC area is 0.978. These high indicators are near to those in Table II.

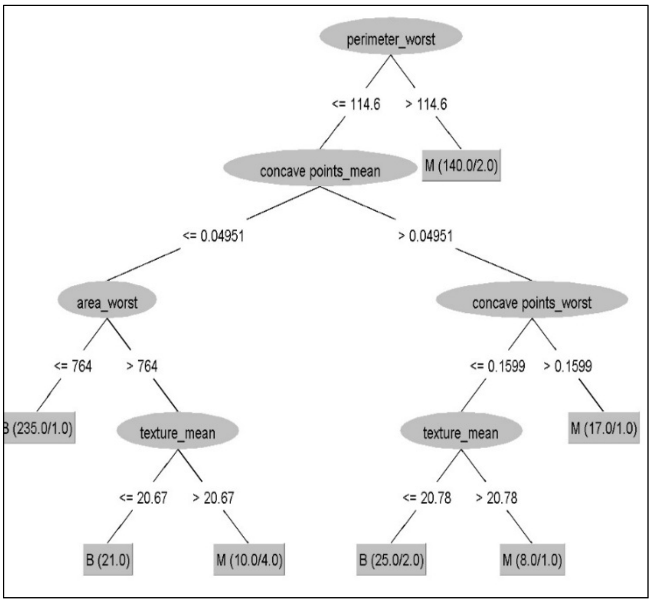


Fig. 3. The slightly pruned "Decision Tree" for the reduced data set; the rectangles depict the leaves, ovals - the tree nodes

Fig.3 presents a somewhat pruned "Decision Tree." This "Tree" has size 13 and comprises seven leaves. Note that the unpruned tree is more branched than the unpruned one: its size is 23 with 12 leaves.

Note that only one attribute ("perimeter_worst") allows separating at once 140 from 212 (it is about two-thirds) patients with malignant tumors. One can see, however, that two patients with benign tumors are also comprised in this leaf. Indeed, one must accept that it is a straight payment for pruning and simplifying the "Decision tree."

Indeed, this pruned "Decision tree" holds only five of eleven attributes of the reduced data set shown in Fig.2. Obviously, it is also the result of the "tree pruning" process. Thus, we got the five-attributed reduced data set, equal by the capacity to the outcome [5], but following a different path. Besides, we can present our data set and its attributes explicitly. The six "Decision Tree" nodes in Fig.3 display those five attributes.

IV. Discussion and conclusions

A. Discussion

WBCD seems too bulky and excessive due to the number of attributes. Recent studies [5] and this report convince us that WBCD can be reduced a few times for the number of attributes with feature selection methods within Machine Learning. Moreover, It is possible without losing the diagnosis precision or with minimal loss.

The number of pertinent attributes may be reduced to five [5], confirmed by the "Decision Tree" in Fig.3 above. These most relevant attributes were, in our case, the following:

- "perimeter_worst (0.685)," "area_worst (0.669)," and "concave points_worst (0.648) ;"
- "concave points_mean (0.635) and "texture_mean (0.159),

The ranks of the most relevant features, according to Fig.2, are pointed out in brackets. One can notice certain, though not ideal, correlations among attributes' ranks and the "decision Tree" nodes' locations.

One can find the optimal classifier concerning this shortest dataset with Auto-Weka. It turned out that "weka.classifiers.meta.Bagging" allows a small number of improved results reported above for the J48 classifier. For example, this classifier gives 0.994 for ROC area and 0.970 for precision and recall.

The association rules, based on one or two attributes, permit useful, though preliminary, conclusions for clinicians too. Besides, the reduced dataset manifests good clusters-to-classes matching. However, the optimal classifiers for these reduced datasets can differ from the best classifiers for complete datasets.

Let us pay special attention to the rising of the "worst" type attributes presence: from 33 % in the complete WBCD to 60 % in the reduced five-attributed one. "Worst" attributes are factually outliers from a statistics point of view; they spoil the sample. On the other hand, they were the high-ranked relevant attributes in our study. It can mean the outliers within biomedical data claim special care. Besides, it turned out they often hold valuable information about a patient.

B. Conclusions

As a result of the investigations mentioned above, it can be concluded that Weka is one of the most suitable software, especially for biomedical data analysis. This is mainly due to the well-designed visualization tools inherent in this software. The visual analysis opportunities, like those offered by Fig.3 above, are a pragmatic necessity for clinicians primarily unfamiliar with the specifics of ML and the associated mathematics.

Reduction, sometimes in a few time, of biomedical data sets, search for association rules, their visualization, infographics like "decision trees" or other similar solutions look like the right path from ML achievements to clinic decision-making. In particular, this conclusion turned out correct for BC diagnostics using WBCD.

We have virtually reduced bulky WBCD to the five-attribute known data set. The reduced dataset is accompanied by clinically acceptable infographics, manifesting the leading impact of "outlying" attributes. Pay attention that these results were attained without essentially losing forecast efficacy.

References

[1] World Bank. 2018. Breast Cancer in Ukraine: The Continuum of Care and Implications for Action. © World Bank, Washington, DC. http://hdl.handle.net/10986/30144 License: CC BY 4.0." URI http://hdl.handle.net/10986/30144 [Accessed: June. 27, 2023].

[2] N. Houssami, G. Kirkpatrick-Jones, N.Noguchi & C. I. Lee (2019) Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice,

Expert Review of Medical Devices. Taylor & Francis, 16(5), pp. 351–362. https://doi.org/10.1080/17434440.2019.1610387

[3] W., Wolberg, O. Mangasarian, N. Street, , and W. Street, (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. [Online]. https://doi.org/10.24432/C5DW2B. [Accessed: June. 27, 2023].

[4] W. T. Mohammad, R. Teete, H. Al-Aaraj, Y. S. Y. Rubbai, et M. M. Arabyat, Diagnosis of Breast Cancer Pathology on the Wisconsin Dataset with the Help of Data Mining Classification and Clustering Techniques, Applied Bionics and Biomechanics [Online serial]. Volume 2022, Article ID 6187275, 9 pages, Available: https://www.researchgate.net/journal/Applied-Bionics-and-Biomechanics-1754-2103 [Accessed June 27, 2023].

[5] S. Kumar and M. Singh, Breast Cancer Detection Based on Feature Selection Using Enhanced Grey Wolf Optimizer and Support Vector Machine Algorithms, Vietnam Journal of Computer Science Vol. 8, No. 2, pp. 177–197, 2021. https://doi.org/10.1142/S219688882150007X

[6] U. Albalawi, S. Manimurugan, and R. Varatharajan, Classification of breast cancer mammogram images using convolution neural network, Concurr. Comput. Pract. Exp., vol. 34, no. 13, 2022. http://dx.doi.org/10.32604/iasc.2021.018607

[7] N. Ahmed, R.- Ibn-Alam, and S. N. Shefat, Performance Evaluation of Data Mining Classification Algorithms for Predicting Breast Cancer, Malaysian J. Sci. Adv. Technol., vol. 2, no. 3, pp. 90–95, 2022. http://dx.doi.org/10.56532/mjsat.v2i3.55

[8] I.H.; Witten, E. Frank,; M. A Hall, C.r J. Pal, (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco (CA). Retrieved 2011-01-19.

[9] R. R. Bouckaert, E. Frank,M. Hall, R. Kirkby, P. Reutemann, A. Seewald, D. Scuse. WEKA Manual for Version 3-9-5. 2020, University of Waikato. URI: https://osdn.net/projects/sfnet_Weka/downloads/documentation/3.9.x/WekaManual-3-9-5.pdf/ [Accessed June 27, 2023].

[10] L. Kotthoff, C. Thornton, and F. Hutter, "User Guide for Auto-WEKA version 2 . 6," 2017, URL: https://www.cs.ubc.ca/labs/algorithms/Projects/autoWeka/manual.pdf [Accessed June 27, 2023].

[11] WEKA Explorer: Visualization, Clustering, Association Rule Mining. May 2023, URL: https://www.softwaretestinghelp.com/Weka-explorer-tutorial/#:~:text=Association%20Rule%20Mining%20Using%20WEKA%20Explorer%201%20Association,the%20Apriori%20algorithm%20for%20learning%20association%20rules.%20 [Accessed June 27, 2023].

[12] Mobashar B., Association Rule Mining with WEKA, 2010, URL: http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/associate.html [Accessed June 27, 2023].

[13] T. Scheffer, "Finding Association Rules That Trade Support Optimally against Confidence." Intelligent Data Analysis, V. 9, Issue 4, pp 381–395, 2005 https://dl.acm.org/doi/10.5555/645805.670142

Machine Learning Methods for Predicting Parkinson's Disease Progression

Yana Teletska
Vasyl' Stus Donetsk National University
Vinnytsia, 21021, Ukraine
y.teletska@donnu.edu.ua

Vira Trofymenko
Vasyl' Stus Donetsk National University
Vinnytsia, 21021, Ukraine
v.trofimenko@donnu.edu.ua
https://orcid.org/0009-0005-8094-227X

Oleh Vietrov
Vasyl' Stus Donetsk National University
Vinnytsia, 21021, Ukraine
o.vietrov@donnu.edu.ua
https://orcid.org/0000-0002-5125-9632

Artem Baiev
Vasyl' Stus Donetsk National University
Vinnytsia, 21021, Ukraine
a.baev@donnu.edu.ua

Abstract **— The article presents the developed machine learning model for the diagnosis and prognosis of a neurological disorder - Parkinson's disease. The choice of the algorithm used to create the model is justified. By employing machine learning methods, particularly Support Vector Machines and k-Nearest Neighbors algorithms, high accuracy and reliability were achieved in predicting disease class labels. The analysis of relationships between clinical, genetic, and biomarker features demonstrated that these characteristics significantly influence the development of Parkinson's disease.**

Keywords — machine learning, prediction of disease progression, Parkinson's disease

I. Introduction

The use of machine learning in medicine is rapidly advancing. Modern technologies enable efficient early diagnosis to determine the most appropriate treatment methods. Neural network construction technologies are applied in disease diagnosis, medical device development, personalized treatment, drug development, and telemedicine. Developed computer products are utilized as reliable tools that aid healthcare professionals in making informed decisions. Medical software programs include automation of medical devices, error detection, prediction, medical image analysis, result interpretation, and test utilization [1].

The economic aspect of using IT technologies in medicine should not be overlooked. According to [2], the global digital healthcare market was estimated at $211.0 billion USD in 2022. As per the forecasts in [2], the market is expected to grow at an average annual rate of 18.6%.

One of the most demanded and advanced applications of artificial intelligence technologies in medicine remains the direction of human disease diagnosis. Figure 1 depicts the distribution of the use of machine learning methods in various diagnostic directions [3].

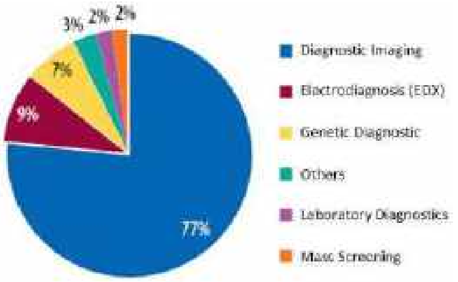


Fig. 1. Distribution of the use of machine learning methods in diagnostics.

In Ukraine, the application of machine learning in medicine is increasingly becoming an effective practical solution. It is worth noting that the implementation of artificial intelligence technologies often stems from private initiatives by entrepreneurs and researchers, frequently in collaboration with state research institutions. An example of this is the medical startup "CheckEye" [4]. Their developed solution allows screening for diabetic retinopathy based on photographs of the eye fundus. The solution was developed in partnership with researchers from the Filatov Institute of Eye Diseases and Tissue Therapy. Another recent Ukrainian development is "Mark" [5]. Utilizing machine learning algorithms, this virtual assistant analyzes test results, health dynamics, medical history, data from health-trackers, and more. Esper Bionics [6] gained global recognition for their startup—a robotic limb prosthesis using advanced machine learning technologies for prosthetic control, providing individuals with improved limb functionality.

Due to the diversity of medical tasks for which artificial intelligence technologies are applied, various machine learning models, such as regression, support vector machines, random forests for supervised learning, and principal component analysis for unsupervised learning, are used to build effective solutions.

The aim of the presented work is to build a model for predicting the progression of Parkinson's disease. Neurological disorders, including Parkinson's disease, are becoming increasingly significant issues for the medical community and society as a whole in the modern world. This disease is one of the complex neurological disorders that have a serious impact on the quality of life of individuals. Among age-related neurological disorders that lead to a range of motor and cognitive symptoms, Parkinson's disease is the second most prevalent. Early initiation of Parkinson's disease treatment is crucial, and for this, reliable early diagnosis of the disease is of utmost importance. The absence of an accurate and reliable diagnostic method for Parkinson's disease can hinder timely treatment initiation and improvement of patients' condition.

The research on the use of machine learning methods for diagnosing Parkinson's disease covers a wide range of scientific literature, as evidenced by a comprehensive review [7]. Additionally, it is worth mentioning contemporary scientific articles dedicated to this topic, which have been published between 2021 and 2023 and were not included in the aforementioned review [7]. These works mainly focus on

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

analyzing the significance of specific disease features for the effective diagnosis of Parkinson's disease at its early stages.

In project [8], the objective is to detect Parkinson's disease using various types of machine learning and deep learning models, such as Support Vector Machine, Random Forest, Decision Tree, k-Nearest Neighbors, and Multi-Layer Perceptron. As a criterion for distinguishing between healthy and diseased individuals, specific features of the voice signal were selected. Works [9-10] are dedicated to various aspects of diagnosing patients based on the characteristics of their voice signals.

The University of Pennsylvania Smell Identification Test (UPSIT) is used to assess the presence of hyposmia (reduced sense of smell) in patients, which is a common symptom of Parkinson's disease. The goal of the study [11] was to create a shortened version of the Italian-adapted UPSIT test. Using several one-dimensional and statistical approaches, 8 items were selected as the most informative, and a model trained on these 8 items performed better compared to the full version with 40 items.

In the work [12], the authors evaluated blood samples using various analytical instruments, such as gas chromatography-MS, capillary electrophoresis-MS and liquid chromatography-MS.

The works [13-14] discuss the most important clinical biomarkers of Parkinson's disease, elucidating their physiological role and functions in this condition. Innovative aptasensors for detecting Parkinson's disease biomarkers using electrochemical methods are introduced for the first time, along with the mention of future alternatives, including ideal analytical platforms for diagnostics in medical care settings.

Research [15] is dedicated to using machine learning algorithms to analyze the length, speed, and width of the step, as well as the variability of step width in patients diagnosed with Parkinson's disease.

One specific research direction is the study of the feasibility of structuring, implementing, validating, and adopting a software tremor simulator capable of generating data related to movement disorders, both for healthy individuals and for pathological conditions, based on raw inertial measurements. This simulator outputs tremor acceleration and angular velocity. The work presented in reference [16] focuses on a particular case, specifically, tremor associated with Parkinson's diseas.

The presented work is conceptually similar to a thorough investigation conducted in [17], where a hypothesis-free disease-relevant network was constructed, focusing on various diseases. Additionally, in [16] investigated drug-gene interactions was studied.

II. Machine Learning Model Creation

To create a machine learning model aimed at predicting the progression of Parkinson's disease, the dataset used was PPMI (Parkinson's Progression Markers Initiative) [18]. PPMI is a dataset specifically established to identify markers of Parkinson's disease progression. This dataset is the result of collaborative efforts by researchers and medical specialists to collect and analyze data related to Parkinson's disease, aiming to understand its progression and develop new diagnostic and treatment methods. It encompasses diverse data types, including clinical data, genetic data, and biomarker data. Clinical data includes information about patients, such as age, gender, symptoms, results of clinical tests, and disease severity scales. These data help establish connections between symptoms and disease progression, as well as investigate risk factors and prognostic indicators. Genetic data reflects genetic markers associated with Parkinson's disease. Researching these markers allows the identification of genetic factors influencing disease risk and progression. Biomarker data consists of information about levels of specific substances in the body that can serve as indicators of the disease's status. Biomarkers can be utilized for diagnosis, predicting disease progression, and assessing treatment effectiveness. This dataset represents a valuable resource for research in the field of Parkinson's disease. It provides the opportunity to study disease progression, establish relationships between clinical features and genetic factors, and explore potential biomarkers for diagnosis and disease monitoring. Such data can contribute to advancing the understanding of Parkinson's disease, developing novel diagnostic and treatment methods, and promoting a personalized approach to managing patients with Parkinson's disease.

The selected dataset consisted of a set of data from 549 patients, categorized as follows: 379 diagnosed with Parkinson's disease, 158 healthy patients, and 52 patients with other symptoms. The distribution of patients by age is presented in Figure 2.

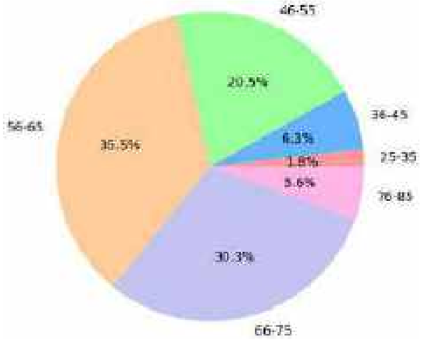


Fig. 2. Distribution of patients by age.

The data were divided into three groups:

1) Genotype:
   - APOE Genotype
   - Number of e4 alleles in APOE Gene
2) Clinical:
   - Age
   - Investigator Diagnosis of Cognitive
3) Biomarkers:
   - Ratio of CSF A-beta 1-42 to CSF Alpha-synuclein (2016 assay)
   - CSF A-beta 1-42 (2016 assay)
   - CSF Alpha-synuclein (2016 assay)
   - Benton Judgement of Line Orientation Score
   - Left caudate
   - Right caudate
   - Contralateral caudate
   - Contralateral count density ratio
   - Contralateral putamen
   - Contralateral striatum

To examine the relationships between variables in the dataset, a correlation matrix was considered (Figure 3).

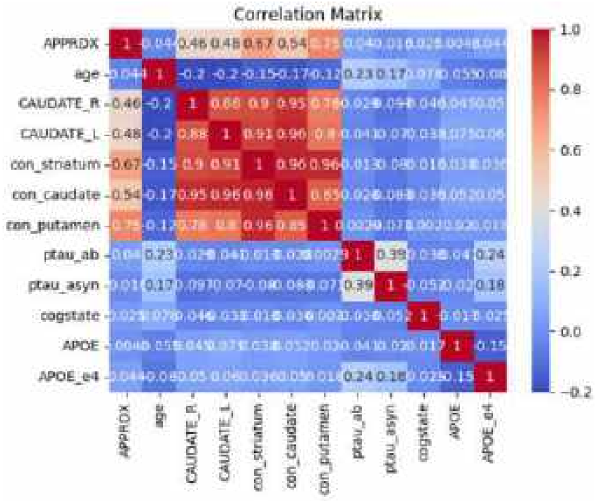


Fig. 3. Correlation Matrix

Using the correlation matrix, one can track the degree of linear dependence between variables and identify whether there is a statistically significant relationship between them. The color scale in the correlation matrix visualizes the degree of correlation between variables. Different shades of color represent different values of the correlation coefficient. This color scale allows for a visual assessment of the degree of dependence between variables. More saturated colors (e.g., red or blue) indicate a stronger correlation between the respective variables. A more saturated red color indicates a positive correlation, while a more saturated blue color indicates a negative correlation.

In addition to correlation analysis, the Random Forest method, using the scikit-learn library, was employed to identify the most important and informative variables in the context of the model:

age = 0.9220219053378521
CAUDATE_R = 0.0769216349161753
CAUDATE_L = 0.06976205675930829
con_striatum = 0.2592766516343571
con_caudate = 0.10859836137307446
con_putamen = 0.2977342630221475
ptau_ab = 0.03531419050428856
ptay_asym = 0.041036940282410256
cjgstate = 0.002021761982299179
APOE = 0.008278892763704143
APOE_e4 = 0.00885305622845005

They are determined based on how well a variable contributes to the reduction of the mean squared error (MSE) or the Gini impurity during the tree splitting in the Random Forest method.

After preprocessing the dataset, it was split into training and testing sets in a ratio of 70:30. 70% of the data was used for training the model, and 30% for testing. For training the initial models, the input variables X were selected:

- 'ptau_asyn': The feature indicates the level of phosphorylated alpha-synuclein in the brain. Alpha-synuclein is a protein that plays a key role in the formation of neurofibrillary tangles, which are characteristic of Parkinson's disease. Higher values of 'ptau_asyn' may indicate the progression of the disease.
- 'ptau_ab': feature indicates the level of phosphorylated amyloid beta-peptide in the brain.
- 'APOE_e4': feature indicates the presence of the ε4 allele of the apolipoprotein E gene.
- 'con_caudate': feature reflects the connectivity or structural connectivity in the caudate nucleus region.

Variable Y (APPRDX) will be the output of the model as it represents the categorization into specific groups: 1 - has Parkinson's symptoms, 2 - healthy, 3 - other disease.

Five models were trained on the variables mentioned above, each using a different algorithm:

- Model1 – Random Forest algorithm;
- Model2 - Decision Tree algorithm;
- Model3 - Gradient Boosting algorithm;
- Model4 - Support Vector Machine algorithm;
- Model5 - k-Nearest Neighbors algorithm.

To determine their effectiveness, the evaluation was performed using five metrics:

- Accuracy: Measures the overall correctness of the model by calculating the ratio of correctly classified examples to the total number of examples.
- Sensitivity: Measures the model's ability to correctly identify positive examples. It is calculated as the ratio of correctly classified positive examples to the total number of positive examples.
- Specificity: Measures the ability of the model to correctly identify negative examples. It is calculated as the ratio of true negative examples to the total number of negative examples.
- F1-Score: combines sensitivity and specificity into a single metric that takes into account both the precision and recall of the model. It is the harmonic mean between precision and recall.
- Receiver Operating Characteristic Area Under the Curve (ROC AUC): easures the model's ability to distinguish between positive and negative classes.

The results of model evaluations can be seen in Table 1.

TABLE I. RESULTS OF MODEL EVALUATIONS

Model	Input data	Accuracy	Sensitivity	Specificity
Model1	'ptau_asyn', 'ptau_ab', 'APOE_e4', 'con_caudate'	0.79	0.54	0.92
Model2		0.73	0.61	0.89
Model3		0.80	0.58	0.94
Model4		0.83	0.58	0.94
Model5		0.76	0.59	0.94

Model	Input data	F1-Score	ROC AUC
Model1	'ptau_asyn', 'ptau_ab', 'APOE_e4', 'con_caudate'	0.63	0.72
Model2		0.66	0.69
Model3		0.68	0.73
Model4		0.68	0.76
Model5		0.68	0.73

For training the following models, the following input variables X were selected:

- 'con_striatum' variable indicates the connectivity between different brain regions, including the striatum. Changes in striatum connectivity may reflect dysfunction in the basal ganglia system, which is characteristic of neurodegenerative diseases. The striatum plays a crucial role in regulating movements, and its connections with other brain regions may be disrupted in Parkinson's patients.

- 'con_putamen' variable indicates the connectivity between different brain regions, including the putamen. The putamen is part of the basal ganglia, which plays an important role in controlling movements and coordination. Changes in the putamen's connections may indicate dysfunction in this structure, which can affect motor symptoms and other manifestations of Parkinson's disease.
- 'APOE' variable is related to the Apolipoprotein E gene, which has significant importance for lipid and cholesterol metabolism in the body. Genetic variations in the APOE gene may influence the risk of developing the disease.

The variable Y – APPRDX. The above-mentioned variables were used to train five models, each using a different algorithm:

- Model1 – k-nearest neighbors algorithm;
- Model2 - decision tree algorithm;
- Model3 - gradient boosting algorithm;
- Model4 - random forest algorithm.

The results of model evaluations can be seen in Table 2.

TABLE II. RESULTS OF MODEL EVALUATIONS

Model	Input data	Accuracy	Sensitivity	Specificity
Model1	'con_striatum', 'con_putamen', 'APOE',	0.91	0.90	1.0
Model2		0.85	0.74	1.0
Model3		0.86	0.75	1.0
Model4		0.91	0.81	1.0

Модель	Input data	F1-Score	ROC AUC
Model1	'con_striatum', 'con_putamen', 'APOE',	0.95	0.94
Model2		0.85	0.88
Model3		0.86	0.88
Model4		0.90	0.90

The two best-performing models were Model4, developed based on the Support Vector Machine algorithm, and Model1, developed based on the k-Nearest Neighbors algorithm. They demonstrated high accuracy. The visualization and performance evaluation of the classification models can be seen in Figure 4. The closer the curve is to the upper left corner of the graph, the better the model performs.

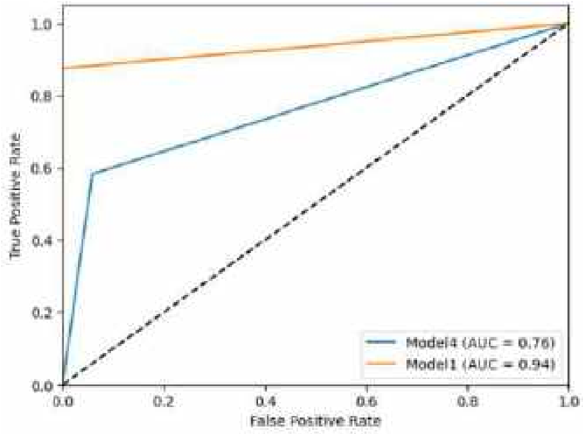


Fig. 4. Dependence between sensitivity (True Positive Rate) and specificity (False Positive Rate) of the models

To enhance accuracy and stability, it was decided to use an ensemble of models, combining different algorithms. This strategy is based on the fact that each individual model has its strengths and weaknesses, and their combination can lead to improved predictive accuracy. In this case, the performance metrics are as follows: Accuracy = 0.96, Sensitivity = 0.88, Specificity = 1.0, F1-Score = 0.93, ROC AUC = 0.94. The results of this ensemble of models can be observed in Figure 5, where the improvement in prediction accuracy is tracked.

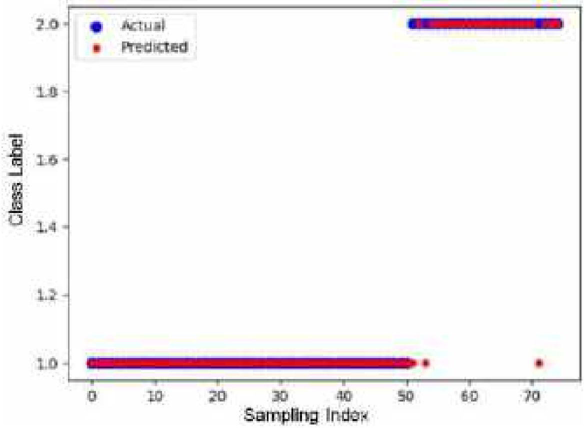


Fig. 5. Visualization of the forecast results of the ensemble of models

The predicted values are represented by the red dots, while the actual values are represented by the blue dots. There is a close alignment between them, indicating that the model successfully predicts class labels with high accuracy. Combining different algorithms in an ensemble allows leveraging their strengths and compensating for their weaknesses. This helps improve the overall predictive accuracy of the model. When multiple models with different algorithms are combined, they can complement each other and reduce the impact of random errors or overlearning, leading to more stable and reliable predictions. The application of model ensembles is a powerful tool for enhancing the accuracy and robustness of predicting the progression of Parkinson's disease. This approach enables obtaining reliable results and instills confidence in the predictive capabilities of the model.

CONCLUSIONS

Parkinson's disease is a chronic neurodegenerative disorder, which is very challenging to study due to the localization of pathology and the changing clinical phenotype over time. Currently, there is no precise diagnostic test for Parkinson's disease. In this regard, the effectiveness of treatment critically depends on early disease detection.

As a result of the research, a successful prognostic model for diagnosing Parkinson's disease was developed. The use of machine learning methods, particularly Support Vector Machines and k-Nearest Neighbors algorithms, enabled achieving high accuracy and reliability in predicting disease class labels. The analysis of interrelations between clinical, genetic, and biomarker features has demonstrated that these attributes significantly influence the development of Parkinson's disease. This reaffirms the importance of integrating various types of data to achieve better diagnosis and understanding of the mechanisms underlying this condition.

To improve the accuracy and stability of the results, it was decided to use an ensemble of models, combining different algorithms. This enhances the reliability of the forecasts and reduces the impact of possible shortcomings in individual models. The obtained research results have significant practical implications. Early diagnosis and identification of high-risk patient groups will enable timely treatment and improve the quality of life for Parkinson's disease patients. Additionally, the knowledge gained about the relationships between features and disease development factors can contribute to further research and the development of new treatment methods.

The research was conducted on a limited dataset; therefore, additional studies using a more extensive dataset could yield more accurate and detailed results. Further research may also involve exploring other machine learning algorithms and investigating new features that may be associated with Parkinson's disease.

Acknowledgment

The work was carried out with the support of the Machine Learning and Data Analytics Laboratory, Faculty of Information and Applied Technologies, Vasyl Stus Donetsk National University..

References

[1] S. Haymond and C. McCudden, “Rise of the Machines: Artificial Intelligence and the Clinical Laboratory, ” J. Appl. Lab. Med., 6(6), 2021, pp. 1640–1654. https://doi.org/10.1093/jalm/jfab075

[2] Digital Health Market Size, Share & Trends Analysis Report By Technology (Healthcare Analytics, mHealth, Tele-healthcare, Digital Health Systems), By Component (Software, Hardware, Services), By Region, And Segment Forecasts, 2023-2030.” https://www.grandviewresearch.com/industry-analysis/digital-health-market

[3] A.A. Vysotskyi, O.O. Surikov and S.V. Vasyliuk-Zaitseva, “Development of artificial intelligence in modern medicine,” Ukrainian Medical Journal, 2 (154) – III/IV, 2023, pp. 1–4. https://doi.org/10.32471/umj.1680-3051.154.241221

[4] Check Eye [Online]. Availale: https://check-eye.com/

[5] Mark [Online]. Availale: https://www.mark.health/

[6] Esper Bionics [Online]. Availale: https://esperbionics.com/

[7] J. Mei, C. Desrosiers and J. Frasnelli, “Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature Front,” Aging Neurosci, Sec. Parkinson’s Disease and Aging-related Movement Disorders, Vol. 13, 2021. https://doi.org/10.3389/fnagi.2021.633752

[8] R. Alshammri, G. Alharbi, E. Alharbi and I. Almubark, “Machine learning approaches to identify Parkinson's disease using voice signal features,” Front Artif Intell, 6: 1084001, 2023. https://doi.org/10.3389/frai.2023.1084001

[9] A. Rana, A. Dumka, R. Singh, M. Rashid, N. Ahmad and M. K. Panda, “An Efficient Machine Learning Approach for Diagnosing Parkinson’s Disease by Utilizing Voice Features,” Electronics, 11 (22), 3782, 2022. https://doi.org/10.3390/electronics11223782

[10] I. Karabayir, S. M. Goldman, S. Pappu, O. Akbilgic, “Gradient boosting for Parkinson's disease diagnosis from voice recordings,” BMC Med Inform Decis Mak, 20(1): 228, 2020. https://doi.org/10.1186/s12911-020-01250-7

[11] A. Landolfi, M. Picillo, M. T. Pellecchia, J. Troisi, M. Amboni, P. Barone and R. Erro, “Screening performances of an 8-item UPSIT Italian version in the diagnosis of Parkinson’s disease,” Neurol Sci, Vol. 44, pp. 889–895, 2023. https://doi.org/10.1007/s10072-022-06457-2

[12] D. J. Zhang, C. Xue, V. B. Kolachalama and W. A. Donald, “Interpretable Machine Learning on Metabolomics Data Reveals Biomarkers for Parkinson’s Disease,” ACS Central Science., 9, 5, pp. 1035–1045, 2023. https://doi.org/10.1021/acscentsci.2c01468

[13] P. Sharma, S. K. Pahuja and K. Veer, “A Systematic Review of Machine Learning Based Gait Characteristics in Parkinson’s Disease”, Mini-Reviews in Medicinal Chemistry, Vol. 22, Is. 8, pp. 1216-1229,2022. http://doi.org/10.2174/1389557521666210927151553

[14] E. Mikuła, J. Katrlík and L.R. Rodrigues, “Electrochemical Aptasensors for Parkinson’s Disease Biomarkers Detection,” Current Medicinal Chemistry, Vol. 29, Is. 37, pp. 5795-5814, 2022. http://doi.org/10.2174/0929867329666220520123337

[15] M. I. A.S.N Ferreira, F. A. Barbieri, V. C. Moreno, T. Penedo and J. M. R.S. Tavares, “Machine learning models for Parkinson’s disease detection and stage classification based on spatial-temporal gait parameters,” Gait & Posture Vol. 98, pp. 49-55, 2022. https://doi.org/10.1016/j.gaitpost.2022.08.014

[16] C. Carissimo, G. Cerro, L. Ferrigno, G. Golluccio and A. Marino, “Development and Assessment of a Movement Disorder Simulator Based on Inertial Data”, Sensors, 22 (17), 6341, 2022. https://doi.org/10.3390/s22176341

[17] M. B. Makarious, H. L. Leonard, D. Vitale et al., “Multi-modality machine learning predicting Parkinson’s disease,” npj Parkinsons Dis, 8, 35, 2022. https://doi.org/10.1038/s41531-022-00288-w

[18] Parkinson's Progression Markers Initiative [Online]. Availale: https://www.ppmi-info.org/

Real-time Classification, Localization and Tracking System (Based on Rhythmic Gymnastics)

Anastasiia Neskorodieva
Vasyl' Stus Donetsk National University
Vinnytsia, 21021, Ukraine
neskorodieva.a@donnu.edu.ua

Maksym Strutovskyi
Vasyl' Stus Donetsk National University
Vinnytsia, 21021, Ukraine
strutovskyi.m@donnu.edu.ua

Artem Baiev
Vasyl' Stus Donetsk National University
Vinnytsia, 21021, Ukraine
a.baev@donnu.edu.ua

Oleh Vietrov
Vasyl' Stus Donetsk National University
Vinnytsia, 21021, Ukraine
o.vietrov@donnu.edu.ua
https://orcid.org/0000-0002-5125-9632

Abstract — **Rhythmic gymnastics, which includes both fast movements and turns of various body parts, as well as a large set of complex non-standard poses and their combinations, was chosen to research the problems of developing an effective real-time classification, localization and tracking system. The choice of rhythmic gymnastics as a model task is dictated by the fact that researchers can create a controlled environment for conducting experiments without restriction of movement and the possibility of rapid repetition of poses, which significantly increases the effectiveness of research and accelerates its implementation.**

Keywords — machine learning, artificial intelligence in sport, rhythmic gymnastics, real-time tracking system

I. Introduction

Based on analysis of modern publications, we can draw an unequivocal conclusion, that in recent years, sport has become one of the most popular fields for machine learning applications and intelligent data analysis. On the one hand, it is due to the diversity of the subject areas and on the other hand, to the commercial prospects of using the obtained models and software products [1-5].

In addition to professionals (athletes, coaches, sports analysts, etc.), fundamental shifts in the sports industry caused by the use of artificial intelligence technologies have become obvious to the regular user as well [6].

From a historical point of view, the first commercial attempts to apply machine learning methods in sports were attempts to predict the results of sports competitions. It is clear that in this case there is an imitation of the ancient tradition of using the methods of mathematical statistics to predict sports results and training efficiency. A similar practice remains relevant even today [7-10].

In addition to predicting the results of sports competitions, the machine learning and intelligent data analysis are used in modern sport to develop optimal tactics and strategy of the game, to increase the effectiveness of training athletes, to organize objective judging of competitions, etc. The analysis of publications shows that among all sport competitions, football and its diverse aspects are the most often the object of research [11-21]. Similar studies on other competitions are less common [22-24].

In order to deepen the study of the problem of human pose recognition, in particular, complex cases of poses, the authors chose rhythmic gymnastics, a sport that has recently become the subject of research on artificial intelligence problems [25-26].

Rhythmic gymnastics, which includes both fast movements and turns of various parts of the body, as well as a large set of complex non-standard poses and their combinations, was chosen to research this issue. An additional advantage of choosing rhythmic gymnastics as a model task is the fact that researchers can create a controlled environment for conducting experiments without movement restriction and the possibility of rapid repetition of poses, which significantly increases the effectiveness of the study and accelerates its implementation.

The presented work is part of the research program implemented by specialists of the interfaculty laboratory of machine learning and intellectual data analysis on the Faculty of Information and Applied Technologies of Vasyl' Stus Donetsk National University.

II. Main Task of Recognizing Human Poses

The pose estimation task, as one of the computer vision tasks, includes detection, association and tracking of semantic key points. From an applied point of view, this is related to the technical task of localizing human joints on image or video to form a skeleton representation.

Automatic human pose detection is a complex problem because it depends image scale and resolution, illumination changes, background obstacles, clothing changes, surroundings, and human interaction with the environment [27]. The list of effective software applications on this topic is significantly limited.

Semantic tracking of key points in video materials in real time requires large computing resources, which significantly limits the accuracy of pose estimation. Conventional pose tracking methods are not fast enough and not occlusion-resistant enough to be effective in implementing applied tasks.

Tracking a person's posture in real-time will allow computers to more subtly and naturally understand human behavior, for example, in autonomous driving, because of

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

detection and tracking of human postures in real-time, algorithms are able to predict much better pedestrians behavior, which provides a more safe driving experience.

Two-dimensional human pose estimation is used to estimate the two-dimensional position or spatial location of key points of the human body from visual effects such as images and videos. Traditional two-dimensional human pose estimation methods use different manual feature extraction methods for individual body parts. Some popular methods include OpenPose, CPN, AlphaPose, and HRNet.

Three-dimensional human posture estimation is used to predict the location of body joints in three-dimensional space. Some methods also reconstruct a three-dimensional human mesh from videos or static images. This approach makes it possible to obtain detailed three-dimensional information about the structure of the human body.

From a technical point of view, 3D estimation of human posture can be performed on monocular images or videos. Although two-dimensional datasets can be easily obtained, collecting accurate three-dimensional pose image annotations is time-consuming and manual labeling is expensive and impractical. Three-dimensional pose estimation had made progress in recent years, like a popular library OpenPose [28].

Depending on the input data size requirements, the pose estimation task can be divided into cases of 2D and 3D estimation. Two-dimensional pose estimation is a prediction of the location of body joints in an image. As a final result, three-dimensional human pose estimation assumes 3D location of all body joints.

Two-dimensional human pose estimation consists in estimating the location of key points in two-dimensional space relative to the image frame. For each key point the model estimates two-dimensional coordinates. It becomes possible to convert the described object into a three-dimensional object by adding the Z dimension to the prediction. Three-dimensional estimation of human pose is a challenging task, even considering the current level of development of machine learning technologies.

Most three-dimensional human pose estimation models involve finding a two-dimensional pose and then trying to synthesize a 3D pose from it. There are also some methods that predict 3D poses without first finding a 2D pose.

The location of human body parts is used to construct a representation of the human body (eg, the pose of the body skeleton) based on the visual input. As a rule, the model approach is used to describe and infer the posture of the human body and to visualize modeled poses.

In most methods, the human body is modeled as an object with joints and limbs. Also this object contains information about the body shape and its kinematic structure. In such a case, a rigid N-joint kinematic model is used [29].

A skeleton-based (kinematic) model flexible and intuitive model of the human body includes a set of joint positions and orientations of the limbs to represent the structure of the human body. Models of this type are used to reflect the relationships between different parts of the body. Skeleton-based models are limited in their representation of texture or shape. This model is used in both two-dimensional and three-dimensional human posture estimation techniques due to its flexibility [30].

A contour-based model is used for two-dimensional pose estimation, therefore contour model is also called planar model. Body parts are represented by several rectangles approximating contours of human body. With the help of principal component analysis it becomes possible to capture the full graph of human body and silhouette deformations (for example, Active Shape Model).

A volumetric model used for three-dimensional pose estimation. For example, GHUM and GHUML are deep learning models trained on a high-resolution whole-body scan dataset of over sixty thousand human configurations to model statistical and articulated three-dimensional human posture and body shape.

Multi-frame assessment of human posture requires large computing power. The performance of human joint detectors for real-time pose tracking is poor, whereas for static image processing, human joint detectors perform very well.

The main problems include the handling of motion blur, video defocus, pose occlusion, and the inability to capture temporal dependencies between video frames.

The use of RNNs causes empirical difficulties in modeling spatial contexts, for example, in the case of a pose occlusions. Modern multi-frame human pose estimation frameworks, such as DCPose, use a large number of time signals between video frames to facilitate keypoint detection.

It is worth noting that human posture recognition has made significant progress in recent years. It went from 2D to 3D pose estimation and from single person to multi-person pose estimation [31]

Algorithms for human pose estimation can be distinguished depending on the use of one of several possible approaches. We can distinguish generative and discriminant methods [32]; depending on the method of work, we distinguish top-down and bottom-up methods [33].

Generally, existing research has focused on the problem of pose estimation from single camera data. It should be noted that there are algorithms whose main idea is to use data from multiple viewpoints/cameras. By combining data in this way, it becomes possible to create more accurate poses and better occlusion processing. A significant limitation of such studies is primarily the lack of appropriate datasets.

Bottom-up and top-down algorithms do not have relational constraints on the final result. The algorithm that predicts joint positions from the input image does not have any filter to reject / correct unnatural posture. This can sometimes lead to an incorrect assessment of a person's posture. To deal with this problem, there is a set of post-processing algorithms that reject unnatural human poses.

The output pose is passed through a learning algorithm that scores each pose based on its likelihood. Poses that score below the threshold are ignored in the testing phase.

A common method of predicting joint locations is to construct confidence maps for each joint.

Preprocessing includes two steps: removing the background and creating a bounding box.

Using MPPE algorithm, bounding boxes are created for each person present in the image. Each of them is then separately evaluated for human pose. In the case of 3D human pose estimation, camera calibration is fundamentally necessary to convert the data to standard coordinates.

Preprocessing includes common Computer Vision based functions such as HoG and SIFT. These features are computed explicitly before feeding the input data to the next learning algorithm. Implicit features refer to feature maps based on deep learning, such as results from complex deep CNNs. These feature maps are part of a complete end-to-end trained pipeline.

In traditional object detection, the human body is perceived only as a bounding frame (square). Traditional pose tracking methods are neither fast enough nor reliable enough to be effective for practical applications, especially in the field of sports where athletes' postures are complex and rapidly changing to be accurately captured. Zenia, for example, is an AI-powered yoga app that uses HPE to help you achieve proper posture during your yoga practice. It uses the camera to detect your pose and assess how accurate your pose is – if it's correct, the predicted pose will be displayed in green, just like the image above. If the pose is incorrect, red will replace green. In addition to yoga, HPE has found application in other forms of exercise.

Most current recognition methods rely heavily on appearance information, taking as input the RGB sequence of entire image regions. Despite the effectiveness of using contextual information about a person, such as their appearance and scene characteristics, this method has large errors when actions have similar contextual information. Methods based on the recognition of human posture and taking only a sequence of human skeletons as input have the same fallacies. Because of this, they suffer from inaccurate assessment of posture or the ambiguity of a person's posture without understanding the context. Models that take into account both external context and pose often get a score bias towards appearance and generalize poorly to videos with little context. To solve this problem, it has been proposed to create an integration of these two features that will dynamically combine appearance and pose streams by observing pose features in real time. Depending on the confidence level of the information, the model allows you to decide what scene information will be used in the integration.

After receiving a video clip as input, the proposed model takes as input its external context sequence “A” and the corresponding pose estimation sequence “P”, and then predicts the probabilities of actions in the video according to the target classes “C”. The appearance sequence consists of RGB frames, and the pose sequence consists of human pose frames. The model architecture processes the two sequences via appearance and pose streams, which will transform “A” and “P” into spatiotemporal characteristics of appearance and pose, respectively. Then, the proposed pose flow-driven feature integrator combines these two features for action recognition. As a result, the action features are fed into the classifier, which gives the probabilities of action classes occurring in the video clip.

When developing the solution, the Mimetics, Kinetics400 AND NTU-RGBD datasets were chosen for training and evaluating the accuracy of the models. These datasets contain videos and short clips of specific human activities that have been annotated for each video.

The peculiarity of the NTU-RGBD dataset compared to others is that the videos in it were taken in laboratory conditions and always allow you to find the full skeleton of people in the video, unlike the other two datasets, whose videos were taken from Internet resources such as Youtube and others.

In total, this dataset contains 56880 videos, each of which corresponds to one of 60 classes of human actions.

Mimetics and Kinetics400 datasets were taken in order to take into account occlusions and incomplete vision of human poses in videos close to real ones.

The main feature is that pose estimation on these videos is extremely difficult and accordingly, it can be tested to what extent the integration of pose estimation analysis will improve accuracy on videos where it alone cannot provide high accuracy results.

For the purity of the validation, the Mimetics dataset was chosen as the dataset to evaluate the model predictions, i.e. its videos will not be involved in the model training process.

The pose stream also takes as input “T” frames from a video clip, using the same sampling scheme as in the context stream, thus extracting key points of the human body from the frames through the use of off-the-shelf two-dimention methods.

The flow architecture uses the same ResNet model as the appearance flow, but with some changes to reduce the memory required. In particular, we skip the first two-stage convolutional layer and the next ResNet layer with max-pooling; this allows us to transform and reduce the size of HP and WP by a factor of four, while maintaining the spatial dimension of the output signal as in the external stream. The search for optimal hyper-parameters for training and the augmentation of training data to increase the accuracy of the model remains promising. The model itself can work with limited efficiency even with pictures.

With the use of Kinetics400, Mimetics and NTU-RGBD datasets, an accuracy research of the final system was conducted. The high accuracy of the context stream was shown on Kinetics and NTU-RGBD datasets. On the Mimetics set, the accuracy was much lower. This is mainly due to the fact that models performed better on camera angles and incomplete poses on the Kinetics and NTU-RGBD datasets than on the new video format.

The pose analysis stream showed high accuracy on the NTU-RGBD set. At the same time, it showed low accuracy on the Kinetics and Mimetics datasets. This is because, unlike NTU-RGBD, the other datasets were collected from videos and clips from Youtube, respectively, due to the occlusions of people or the perspective of the camera, the model could not capture the full skeleton of people's limbs in the video.

The result of evaluating a figure in video content is its formalized description in the category space. Such categories are the position of the limbs, head, etc. Each component of the figure is described in its own specific category space. Each j-th shape is described by two objects: A_j – a fragment of a video frame and a vector $B_j = (b_{j1}, b_{j2},...,b_{jN})$, which defines the category of each of the N components of the shape, $j = 1..M$.

As a result of processing a video frame fragment A_j, the recognition system produces the result in the form of a vector $C_j = (c_{j1}, c_{j2},...,c_{jN}$). The meaning of each coordinate of this vector is similar to the coordinates of the vector B_j. The task of optimizing the system is to choose its parameters P that will ensure the minimum deviation between the vectors B_j and C_j over the entire sample of M shapes. With the shape-by-shape approach, the similarity between the vectors B_j and C_j is calculated at the j-th iteration of the learning algorithm and the parameters P are immediately modified to increase this similarity. At the same time, the intensity of the modification should be small enough to make the learning process fast, and small enough to avoid losing the experience from previous examples during the learning process. It is also possible to calculate the inverse indicator – the disjunction between vectors B_j and C_j, and then the optimization will be to minimize the disjunction (which is equivalent to maximizing the similarity). The batch approach first calculates the similarity for each pair of vectors B_j and C_j, $j = 1..M.$, then averages the similarities over the entire sample, and modifies the parameters P based on the averaged metric.

The final model showed that it becomes possible to achieve higher accuracy of video prediction results on datasets by combining pose streams and video context. This allows the strengths of both approaches to be utilized.

For the study [34], the authors created their own dataset consisting of 1000 video performances of female athletes with the objects hoop, ball, mace and ribbon This dataset is designed only for the rules and evaluation system of athletes in 2016-2020.

To study the classification methods, the authors of this article built their own dataset, which contains test and training datasets for further training of the sports elements classification model. The five most popular elements in terms of performance were selected for classification.

The authors put forward the following requirements for the data set:

- the images of individual athletes, which should occupy at least 50% of the image length.
- the image should show all parts of the athlete's body.
- the minimum image quality should be 240 by 360.

The following resources were used to collect the data set:

- video recordings of athletes' performances at the XXXII Olympic Games in Tokyo [35, 36].
- photo reports from the competitions by the official photographer of the National Rhythmic Gymnastics Team of Ukraine Maria Muzychenko [37].
- photo reports by photographer Ihor Sakhatsky [38].
- video of performances [39].

In general, the implementation of the project can be schematically divided into the following stages: 1) Dataset collection 2) Data marking 3) Retraining neural networks on marked data 4) Combining detector models with tracking algorithms. 5) Adding to the obtained result a set of basic algorithms 6) Optimizing the obtained combinations 7) Modeling the neural network architecture 8) Software development 9) Testing the solution.

Acknowledgment

The present research is carried out as a part of the basic research program of Ministry of Education and Science of Ukraine, the project number is 0121U109526.

Conclusions

From the analysis of modern publications, it becomes obvious that certain classes of posture assessment tasks are well researched, and the available technical implementations allow us to single out directions where high accuracy is achieved for the assessment of basic movements (walking, slow hand movements, jumping up, etc.).

The assessment of posture for fast movements and non-standard postures of a complex type due to the large number of limbs involved in movement remains significantly less researched. A large number of synchronized videos with fast movements and sequences of complex poses from different angles allows forming a dataset necessary for further research and implementation of the obtained results both in socially significant industries and in the market of commercial services using artificial intelligence technologies.

Engineers of the laboratory of machine learning and intellectual analysis of the Vasyl’ Stus Donetsk National University develop a computer system that can be used to increase the objectivity of sports refereeing in rhythmic gymnastics competitions, as well as become an alternative to the traditional refereeing system in the case of competitions held in a remote format . Scaling the task, the system can also be used to diagnose problems with the nervous system and musculoskeletal system of a person.

All existing models have certain shortcomings, it was experimentally established that the use of ready-made solutions is insufficient for the implementation of the given task. The main problem of the existing open solutions have the low efficiency of the system for working with a large crowd of people, that is, the problem of combating occlusions and research on improving the quality of object detection models is extremely relevant. Also, the problem of synchronizing data from the camera, for combining visual images from several cameras, due to technical delays in the operation of the device and low bandwidth, turned out to be significant.

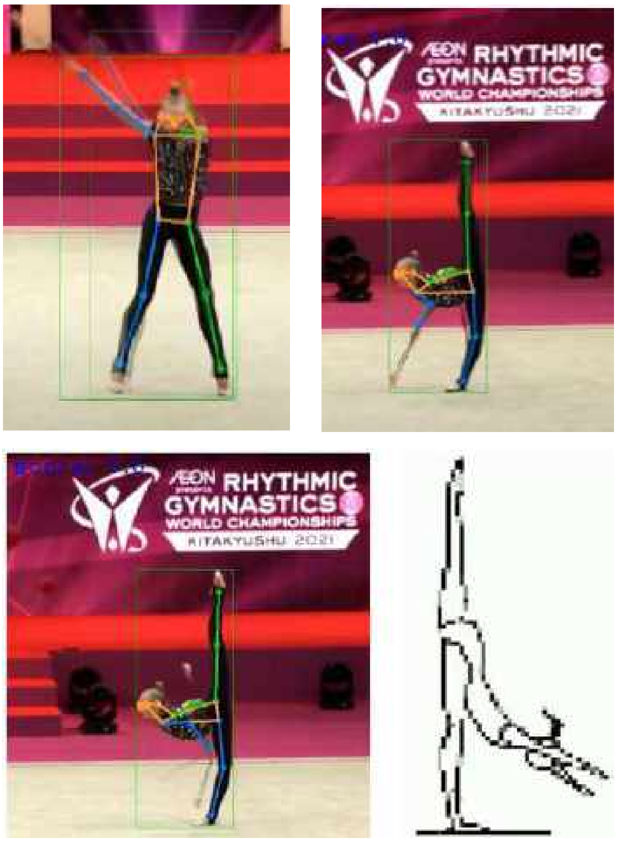


Fig. 1. An example of the program

The main point of the research is the comparison of the quality of dataset formation using motion capture technologies (through a suit and camera) and manual annotation of 2D/3D poses. Further research on data sets generated by different technologies will allow the development of a new training methodology on the data sets to achieve the highest final model accuracy. The presence of synchronized video and ground truth data will provide an accurate metric for assessing the reliability of pose estimation class models.

As a result of the project, a unique dataset of human poses will be created. The dataset differs from existing analogues by a large number of synchronized videos with fast movements and sequences of complex poses from different angles. The practical application of the research results will be much wider than just the use in the sports field.

Based on the built model, it will be possible to expand the use functionality of the security camera system, which can be used to find potentially dangerous positions for others (for example, shooting or hitting) and for the person himself (for example, fainting). A possible application in the future is an application for attention control systems, improving the quality of online learning and workflow.

The actual usage of the suggested complex model is medicine usage. In particular, detecting such postural problems as scoliosis by analyzing the patient's posture disorders, physical therapy and studying the cognitive development of the young children brain by monitoring motor functionality. Improving the functionality of video surveillance to achieve an automatic alarm system in the event of potentially dangerous situations.

Human pose analysis can be methodologically extended for sign language detection and voice translation tasks by extending human-computer interaction for gesture control and developing motion capture systems without markers or sensors usage.

Finally, the direct socially significant task, the implementation of which the project is aimed at, is the fundamental possibility of improving the learning process in sports, providing an opportunity to clearly see a specific problem in the execution of a certain sequence of actions, providing an additional tool for analyzing the success of performing movements in the correct manner. This will make it possible to increase the objectivity of refereeing, and in the future will provide an opportunity to hold national and international sports events remotely.

References

[1] A.W. Pearson , "The A.I. Sports Book: How AI and Machine Learning can revolutionize the sports", Independently published, 2019. 422 p.

[2] N. Chmait & H. Westerbeek, "Artificial Intelligence and Machine Learning in Sport Research: An Introduction for Non-data Scientists", Front Sports Act Living, vol. 3, 2021, 682287. https://doi.org/10.3389/fspor.2021.682287

[3] C. Richter, M. O'Reilly & E. Delahunt, "Machine learning in sports science: challenges and opportunities", Sports Biomechanics, 2021. https://doi.org/10.1080/14763141.2021.1910334

[4] U. Brefeld, J. Davis, M. Lames & J.J. Little, "Machine Learning in Sports", Dagstuhl-Seminar, vol. 11, issue 9, 2021, 21411. https://doi.org/10.4230/DagRep.11.9.45

[5] R.M. Musa, Z. Taha, A.P.P.A. Majeed & M.R. Abdullah, "Machine Learning in Sports", Springer Singapore, SpringerBriefs in Applied Sciences and Technology, 2019, 45 p. https://doi.org/10.1007/978-981-13-2592-2

[6] A. Rizzoli, "7 Game-Changing AI Applications in the Sports Industry", https://www.v7labs.com/blog/ai-in-sports

[7] R. Bunker & T. Susnjak, "The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review", Journal of Artificial Intelligence Research, Vol. 73, 2022. https://doi.org/10.1613/jair.1.13509

[8] S. Lotfi & M. Rebbouj, "Machine Learning for sport results prediction using algorithms", International Journal of Information Technology, 3 (3), 2021, pp. 148-155. https://doi.org/10.52502/ijitas.v3i3.114

[9] R.P. Bunkera & F. Thabtah, "A machine learning framework for sport result prediction", Applied Computing and Informatics, vol. 15, issue 1, 2019, pp. 27-33. https://doi.org/10.1016/j.aci.2017.09.005

[10] T. Horvat & J. Job, "The use of machine learning in sport outcome prediction: A review", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 12, issue 2, 2022. https://doi.org/10.1002/widm.1380

[11] A. Rossi, L. Pappalardo & P. Cintia, "A Narrative Review for a Machine Learning Application in Sports: An Example Based on Injury Forecasting in Soccer", Sports, 10(1), 5, 2022. https://doi.org/10.3390/sports10010005

[12] A. Majumdar, R. Bakirov, D. Hodges, S. Scott & Tim Rees, "Machine Learning for Understanding and Predicting Injuries in Football", Sports Med - Open, vol. 8, 2022, 73. https://doi.org/10.1186/s40798-022-00465-4

[13] S.S. Lövdal, R.J.R.D. Hartigh & G. Azzopardi, "Injury Prediction in Competitive Runners With Machine Learning", International Journal of Sports Physiology and Performance, 16, 2021, 1522-1531 https://doi.org/10.1123/ijspp.2020-0518

[14] R.M. Musa, A.P.P.A. Majeed, N.A. Kosni, M.R. Abdullah, "Machine Learning in Team Sports: Performance Analysis and Talent Identification in Beach Soccer & Sepak-takraw", Springer Singapore, SpringerBriefs in Applied Sciences and Technology, 2020, 61 p. https://doi.org/10.1007/978-981-15-3219-1

[15] C. Li, S. Kampakis, P. Treleaven, "Machine Learning Modeling to Evaluate the Value of Football Players". https://doi.org/10.48550/arXiv.2207.11361

[16] H. Li, C. Cui & S. Jiang, "Strategy for improving the football teaching quality by AI and metaverse-empowered in mobile

internet environment", Wireless Netw, 2022. https://doi.org/10.1007/s11276-022-03000-1

[17] X. Zuo, "Visualization of Football Tactics with Deep Learning Models", 2022, https://doi.org/10.1155/2022/9259328

[18] F. Rodriguesa, Â. Pintob, "Prediction of football match results with Machine Learning", Procedia Computer Science, vol. 204, 2022, pp. 463-470 https://doi.org/10.1016/j.procs.2022.08.057

[19] K. Tuyls, S. Omidshafiei, P. Muller and others, "Game Plan: What AI can do for Football, and What Football can do for AI", Journal of Artificial Intelligence Research, vol. 71, 2021, pp. 41-88. https://doi.org/10.1613/jair.1.12505

[20] M. Herold, F. Goes, S. Nopp and others, "Machine learning in men's professional football: Current applications and future directions for improving attacking play", International Journal of Sports Science & Coaching, vol. 14, issue 6, 2019, https://doi.org/10.1177/1747954119879350

[21] G. Anzer, P. Bauer, U. Brefeld, D. Faßmeyer, "Detection of tactical patterns using semi-supervised graph neural networks", 2022. https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/6227709e4d7acb78147f7bcf_Detection%20of%20Tactical%20Patterns%202.pdf

[22] S. Wilkens, "Sports prediction and betting models in the machine learning age: The case of tennis", Journal of Sports Analytics, vol. 7, no. 2, 2021, pp. 99-117. https://doi.org/10.3233/JSA-200463

[23] J. Rothschild, "Predicting daily recovery during long-term endurance training using machine learning analysis", Preprint, https://doi.org/10.51224/SRXIV.191

[24] S.K. Andrews, K.L. Narayanan, K. Balasubadra & M.S. Josephine, "Analysis on Sports Data Match Result Prediction Using Machine Learning Libraries", Journal of Physics: Conference Series, 1964, 2021, 042085. https://doi.org/10.1088/1742-6596/1964/4/042085

[25] M. Mack, M. Bryan, G. Heyer & T. Heinen, "Modeling Judges' Scores in Artistic Gymnastics", The Open Sports Sciences Journal, 12 (1), 2019, pp.1-9. https://doi.org/10.2174/1875399X01912010001

[26] M. Pino Díaz-Pereira, I. Gómez-Conde, M. Escalona & D.N. Olivieri, "Automatic recognition and scoring of olympic rhythmic gymnastic movements", Human Movement Science, vol. 34, 2014, pp. 63-80. https://doi.org/10.1016/j.humov.2014.01.001

[27] L. Pishchulin, "Poselet conditioned pictorial structures", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 588-595 https://doi.org/10.1109/CVPR.2013.82

[28] A. Toshev, C. Szegedy, "Deeppose: Human pose estimation via deep neural networks", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653-1660. https://doi.org/10.1109/CVPR.2014.214

[29] Z. Cao, T. Simon, Shih-En W. Yaser, Sheikh "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields" https://arxiv.org/pdf/1611.08050.pdf

[30] Y. Du, W. Wang, L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 7, No. 12, 2015, pp. 1110-1118. https://doi.org/10.1109/CVPR.2015.7298714

[31] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation", Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2334-2343, 2017. https://doi.org/10.1109/ICCV.2017.256

[32] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2988. https://doi.org/10.1109/ICCV.2017.322.

[33] R. Girshick, "Fast R-CNN", https://arxiv.org/pdf/1504.08083I

[34] L.-A. Zeng, F.-T. Hong, W.-Sh. Zheng, Q.-Zh.Yu, W. Zeng, Y.-W. Wang & J.-H. Lai, "Hybrid Dynamic-static Context-aware Attention Network for Action Assessment in Long Videos", https://arxiv.org/abs/2008.05977, 2020, pp. 1-10.

[35] Olympics Gymnastics: Rhythmic Gymnastics - Individual All-Around-Qualification 1&2 | Tokyo 2020. https://www.youtube.com/watch?v=uRzmkLF8MVI (date of access: 27.05.2023).

[36] Olympics: FULL Rhythmic Gymnastics Individual All Around Final at Tokyo 2020. https://www.youtube.com/watch?v=v6ZuroWdLTs (date of access: 27.05.2023).

[37] https://muzychenko.photos/our-services/sports-photography (date of access: 27.05.2023).

[38] https://sakhatskyi.com/portfolio/ (date of access: 27.05.2023).

[39] Ukrainian RG Federation: Viktoriia Onopriienko Ball Qual 26,200 - World Championships Kitakyushu 2021. https://www.youtube.com/watch?v=IKzuWUIe8Rc (date of access: 27.05.2023).

Features of Using the Prophet Package for Forecasting the Local Weather Situation

Yuriy Korchak
Department of electronics and computer technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
yuriy.korchak@lnu.edu.ua

Bohdan Ivashko
Department of electronics and computer technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
bohdan.ivashko.feim@lnu.edu.ua

Yuriy Furgala
Department of electronics and computer technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
furgala@mail.lviv.ua

Abstract — **The paper analyzes the features of the Prophet library as a platform for forecasting the local meteorological situation. Annual and weekly forecasting of temperature changes in the environment in the city of Lviv was made, both without using and using the influence of exogenous parameters. Almost identical results were obtained in these two cases. The calculation of absolute and relative forecasting errors indicates a good quality of forecasting temperature changes using the Prophet package, at least for a weekly time frame.**

Keywords* — *time series, mathematical model, Prophet library, exogenous parameter, absolute error, relative error, forecasting quality.

I. INTRODUCTION

Forecasting is a fairly popular and probably the most common analytical problem that arises when working with time series. However, it is not easy to get reliable forecasts - this requires serious training of a specialist who solves such a problem, as well as the availability of appropriate and easy-to-use software. To forecast time series in the *Python* and *R* programming languages, various approaches based on the *ARIMA* package are most often used - models of autoregressive integration of the moving average [1-5]. However, in this case, it is necessary to overcome certain difficulties associated with the selection of appropriate own setting parameters (the number of elements of the moving average, the order of autoregression, the number of differences to ensure the stationarity of the time series, etc.), which is a rather time-consuming process [6, 7]. To simplify the procedure of all these processes, in February 2017, *Core Data Science* employees from *Facebook* released a new library for working with time series - *Prophet* [8], which was initially used to forecast a large number of business indicators. This application software package is largely devoid of the above-mentioned shortcomings and allows you to create accurate predictive models in (semi-)automatic mode. This work examines the features of using this library for forecasting the local weather situation, since, as is known [9], weather observation statistics are presented in the form of time series. In addition, an attempt was made to estimate the time limits of the applicability of the *Prophet* platform in this case.

II. MODEL PROPHET

The *Prophet* package is distributed free of charge under the *MIT* license, an open and free software license developed by the Massachusetts Institute of Technology (USA). The methodology of this library is based on the procedure for fitting additive regression models (*GAM* – Generalized Additive Models) of this type [8, 10]:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t ,$$

where $g(t)$ and $s(t)$ are functions that approximate the series trend and seasonal fluctuations (for example, annual, weekly, etc.), respectively, $h(t)$ is a function that reflects the effects of holidays and other influential events, ε_t – normally distributed random disturbances.

To approximate the listed functions, such methods as trend, annual and weekly seasonality, holidays and weekends or other days, during which the properties of the time series can change significantly, for example, certain natural phenomena, are used [10]. Let's consider these methods in more detail.

The trend of the series $g(t)$ is described by piecewise linear or logistic growth. The advantage of applying such approaches to modeling is the simplicity and comprehensibility of the results, the speed of learning and the creation of a forecast.

Recall that the piecewise linear regression function is described by the expression

$$g(t) = at + b + c_1\left|t - t_1\right| + c_2\left|t - t_2\right| + ... + c_n\left|t - t_n\right|$$

and is characterized by a linear dependence within certain time intervals (-∞; t_1), (t_1; t_2),…, (t_n; +∞). Here, the coefficients c_i and a can be expressed in terms of the angular coefficients of the slope k_i of the straight lines at separate intervals:

$$c_i = \frac{k_i - k_{i-1}}{2} (i = 1, \quad 2, \quad ..., \quad n); \quad a = \frac{k_0 + k_n}{2} .$$

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

A continuous piecewise linear function is often also called a linear spline.

Logistic function in the form

$$g(t) = \frac{C}{1 + \exp(-k(t - b))}$$

allows simulating growth with saturation, when in the case of an increase in the indicator (b), its growth rate (k) decreases. A typical example is the growth of the audience of an application or website. In addition, the library is able to automatically select optimal trend change points based on historical data. Although they can also be set manually (for example, if the release dates of new functionality are known, which will greatly affect key indicators).

As for the seasonality parameter s(t), the annual seasonality is described by the partial sums of the Fourier series, the number of members (order) of which determines the smoothness of the function, and the weekly seasonality is an indicator variable. Since in many cases the data includes not only quantitative, but also qualitative characteristics (gender, social status, position, settlement, presence of disease, etc.), the very creation of indicator variables or dummy variables allows transforming these qualitative predictors in quantitative In the case of weekly seasonality, 6 additional signs are added, for example [*Monday*, *Tuesday*, *Wednesday*, *Thursday*, *Friday*, *Saturday*], which take the value 0 or 1 depending on the date. The *Sunday* sign, which corresponds to the seventh day of the week, is not added because it will depend linearly on the other days of the week and this will affect the model.

The influence of anomalous days (regular (weekends) and irregular (holidays, days of natural disasters, days of mass sales)), which cause significant changes in the properties of the time series, is described by the function h(t). This approach makes it possible to model, for example, holidays and weekends, giving the possibility to forecast emissions in a series.

The estimation of the parameters of the model to be fitted is performed using the principles of Bayesian statistics (either by the method of finding the maximum a posteriori (*MAP*) or by full Bayesian inference) [11]. For this, the *Stan* probabilistic programming platform is used. The *Prophet* package is actually a convenient interface for working with this platform from the *R* environment. For the *Python* programming language, there is a similar library called *fbprophet*.

III. Results And Discussion

The interface of the *Prophet* library fully corresponds to the popular machine learning library *scikit-learn* [12]: we create a model, train it with the help of the *.fit()* method, and build predictions with *.predict()*. For training, you need to pass a *DataFrame* with columns:

- *ds* – time, the field format must be *datetime*;
- *y* – target variable in numeric format.

To get the prediction, you are required to pass a new *DataFrame* that contains the *ds* column. For this, the library has a *make_future_dataframe* function that accepts the parameter *periods* – the period for which forecasting is performed and *freq* – the frequency of the time series (the default value is a day). After performing these procedures, training was implemented for the last 5.5 years and a forecast for the next year was made.

As a forecast, *Prophe*t returns a *DataFrame* with a large number of columns. The most important of them:

- *ds* – timestamp for the predicted value;
- *yhat* – predicted value;
- *yhat_lower* – the lower limit of the forecast;
- *yhat_upper* – the upper limit of the forecast.

The library provides quite convenient rendering methods. One of them - *Prophet.plot* - displays a prediction graph (Fig. 1). In this case, forecasted temperature values are built for one value per day, namely for 11:00 p.m. As we can see from Fig. 1, the cyclicity and trends of its change are quite well observed in the prediction of the temperature of the external environment in the city of Lviv.

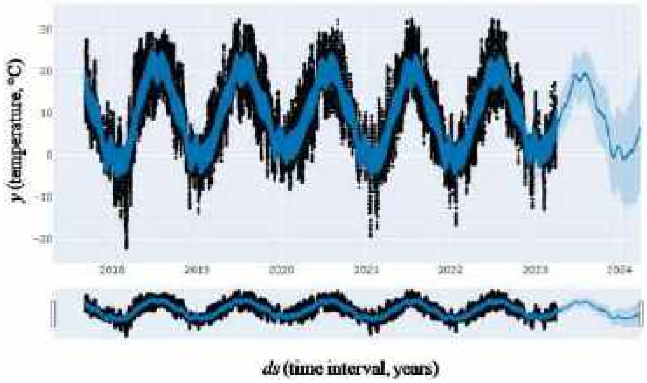


Fig. 1. Application of the *Prophet* model for forecasting temperature changes for the coming year in the city of Lviv. In the figure, the actual temperature values are marked in black, and the predicted values in blue

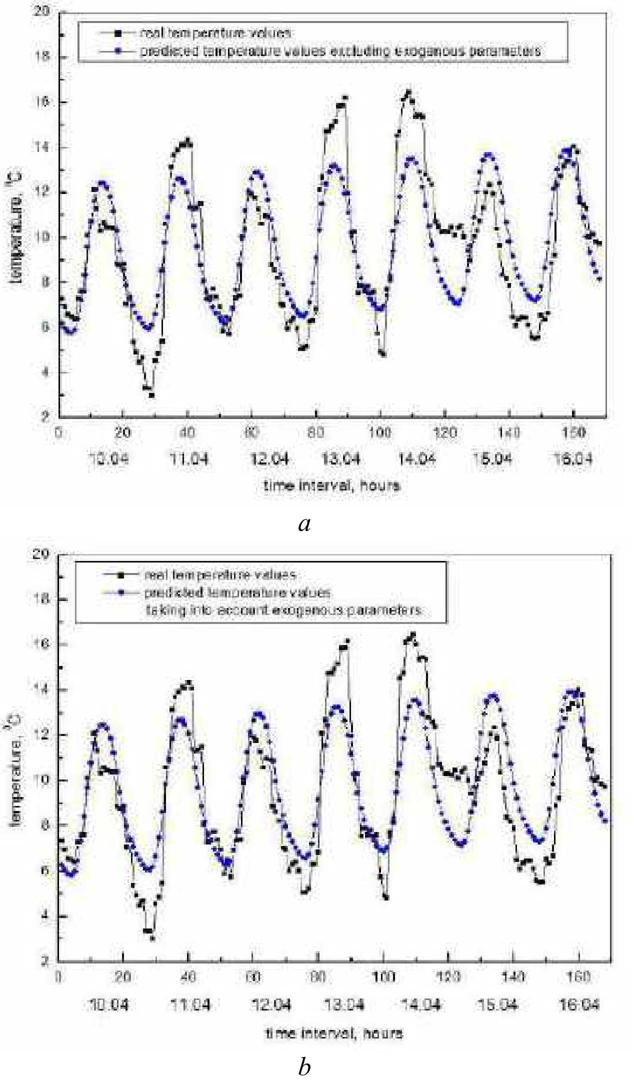
In order to assess the applicability limits of the *Prophet* library, a weekly hourly prediction of the change in the temperature of the external environment in the city of Lviv in the period from 00:00 04/10/2023 until 11:00 p.m. 04/16/2023 was implemented without taking into account (Fig. 2, a) and taking into account (Fig. 2, b) the influence of regressor variables (exogenous parameters). Atmospheric pressure and humidity were considered as exogenous parameters. In this case, their forecast changes were taken into account, which significantly complicated the forecasting process. It is important to note the non-obvious fact that taking into account the influence of exogenous parameters practically does not change the quality of the predictions of temperature changes (see Fig. 2).

Absolute (*MAE*, *MSE*) and relative (*MAPE*, K_T) criteria of forecasting accuracy were considered to assess the quality of the implemented temperature change forecasting, namely:

- $MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$ - mean absolute error;
- $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ - mean squared error;
- $MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$ - mean absolute percentage error;

- $K_T = \dfrac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^2}}$ – Theil's inequality coefficient.

In the given expressions for errors y_i and $\widehat{y}_l$ – real and predicted values of the modulated variable (in our case, temperature), respectively, and n – the number of observations. The first three parameters were calculated using the *performance_metrics()* function.



a

b

Fig. 2. Application of the *Prophet* model to predict the temperature change for the coming week without (a) and with (b) the use of regressor variables. In the figure, the actual temperature values are marked in black, and the predicted values in blue

The value of the main absolute error of *MAE* without taking into account the influence of exogenous parameters was in the range of 0.644-1.556, and with the influence of exogenous parameters in the range of 0.589-1.577. For the mean squared error *MSE*, values in these two cases were obtained, respectively, within the range of 0.497-3.326 and 0.439-3.363. These data indicate a certain expansion of the range of dispersion of the absolute parameters *MAE* and *MSE* in the case of taking into account the influence of regressor variables, although their values fully satisfy the quality of the temperature forecast.

In fig. 3 shows the calculated values of the *MAPE* parameter, presented as a percentage, depending on the number of forecast values during the considered weekly time period excluding and taking into account the influence of exogenous parameters. As can be seen from fig. 3, there is an overlap of the behavior of the *MAPE* parameter in these two cases.

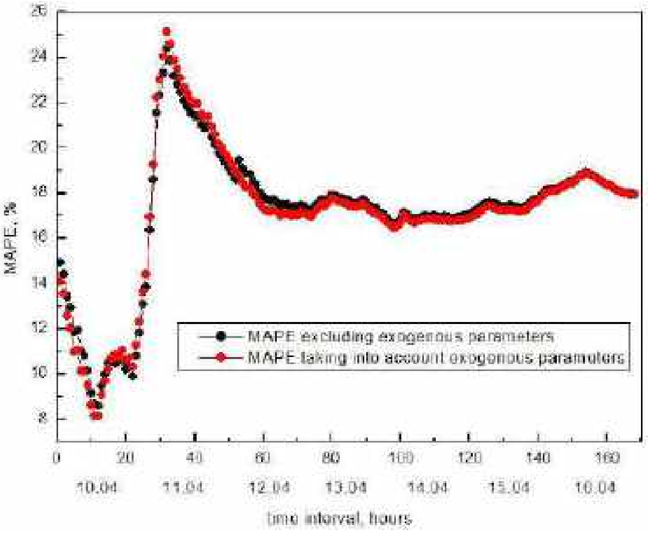


Fig. 3. Dependence of the *MAPE* error on the number of observations during the considered weekly period in the case of using the *Prophet* model for forecasting temperature changes. In the figure, the values of the *MAPE* parameter without taking into account the influence of regressor variables are marked in black, and in red - with their influence taken into account

Analyzing these data and taking into account the generally accepted approaches to assessing the accuracy of time series forecasting [13], it can be concluded that during almost the entire weekly forecasting period, the value of the average absolute specific error of *MAPE* is within 10-20%, which indicates a good quality of forecasting. Although it is worth noting that there are two small time intervals in which this parameter exceeds these limits. In particular, in the time interval from 8:00 a.m. to 3:00 p.m. on April 10, 2023, the accuracy of forecasting is high (MAPE < 10%), and in the time interval 4:00-10:00 p.m. on April 11, 2023, the forecasting accuracy is satisfactory (MAPE 20-40%).

Another important indicator of the accuracy of the model and its compatibility is the Theil's coefficient K_T – a dimensionless value that is between 0 and 1. The closer this parameter is to zero, the better the predictive qualities of the model [14-16]. For the given temperature forecast, the value of the Theil's coefficient without taking into account the influence of exogenous parameters was within the range of 0.0389-0.0927, and with the influence of exogenous parameters within the range of 0.0405-0.0928.

Conclusion

Although the *Prophet* library was initially developed for the implementation of forecasting of economic processes, the data presented in the work indicate a fairly good quality of the implementation of forecasting of the local weather situation. The *Prophet* package is insensitive to missing values, trend biases, and significant outliers, which is an important advantage over *ARIMA*. Another advantage is the rather high speed of learning, as well as the possibility of using large time series [17].

As shown by the assessment of absolute and relative errors of forecasting temperature changes, the application of the *Prophet* library gives a fairly good quality of weekly prediction. An important result of the conducted temperature forecasting is the fact that the lack of consideration of the influence of such exogenous parameters as atmospheric pressure and humidity does not impair the quality of the forecast. Therefore, if it is necessary to speed up forecasting processes, it is quite possible to ignore their influence on temperature as a modulated variable.

REFERENCES

[1] P. Whittle, Hypothesis Testing in Time Series Analysis. Uppsala: Almquist and Wiksells Boktryckeri AB, 1951.

[2] G. P. E. Box and G. M. Jenkins, Time Series Analysis: Forecasting and Control, San Francisco: Holden-Day, 1970.

[3] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, Melbourne: OTexts.org/fpp, 2013.

[4] B. Artley, “Time Series Forecasting with ARIMA, SARIMA and SARIMAX”. [Online]. Available: https://towardsdatascience.com/time-series-forecasting-with-arima-sarima-and-sarimax-ee61099e78f6

[5] William W. S. Wei, Time Series Analysis: Univariate and Multivariate Methods, New York: Addison-Wesley Publishing Company, Inc, 2006.

[6] O. Dzendzelyuk, I. Kostiv and V. Rabyk, “Building ARIMA Time Series Models for Weather Data Predicting Using R Programming Language”, Electronics and information technologies, Issue 3, 2013, pp. 211-219. (in Ukrainian)

[7] Yu. Korchak, Yu. Furgala, Yu. Panasiuk and D Rozhankivskyi, “Application of Adaptive Predicative Analytics for Forecasting the Local Weather Situation”, Electronics and information technologies, Issue 18, 2022, pp. 20-33. (in Ukrainian)

[8] S. J. Taylor and B. Letham, “Forecasting at Scale”, The American Statistician, vol. 72, no 1, 2017, pp. 37–45.

[9] Bryan F.J. Manly, Statistics for Environmental Science and Management, 2nd ed., New York: Chapman and Hall/CRC, 2008.

[10] M. Krieger, “Time Series Analysies with Facebook Prophet: How it works and How to use it”. [Online]. Available: https://towardsdatascience.com/time-series-analysis-with-facebook-prophet-how-it-works-and-how-to-use-it-f15ecf2c0e3a

[11] “Bayesian statistics”. [Online]. Available: https://en.wikiversity.org/wiki/Bayesian_statistics

[12] “Scikit-learn. Machine Learning in Python”. [Online]. Available: https://scikit-learn.org/stable/

[13] “What is a good MAPE score?”. [Online]. Available: https://stephenallwright.com/good-mape-score/

[14] H. Theil, Applied Economic Forecasting, Chicago: Rand McNally & Company, 1966.

[15] F. Bliemel, “Theil’s Forecast Accuracy Coefficient: A Clarification”, Journal of Marketing Research, vol.10 No 4, 1973, pp. 444-446.

[16] Dennis A. Ahiburg, “Forecast evaluation and improvement using theil’s decomposition”, Journal of Forecasting, vol.3 No 3, 1984, pp. 345-451.

[17] T. V. Hnot and M. V. Nehrey, “Data Science Algorithms in Business Processes Modeling”, Economy and society, Issue 12, 2017, pp. 743-751. (in Ukrainian)

Detection of Surface Defects Inside Concrete Pipelines Using Trained Model on JetRacer Kit

Roman Mysiuk
Department of System Design
Ivan Franko National University of Lviv
Lviv, Ukraine
0000-0002-7843-7646

Volodymyr Yuzevych
Department of Electrophysical Methods of Non-Destructive Testing
Karpenko Physico-Mechanical Institute of the NAS of Ukraine
Lviv, Ukraine
0000-0001-5244-1850

Iryna Mysiuk
Department of System Design
Ivan Franko National University of Lviv
Lviv, Ukraine
0000-0002-3641-4518

Yuriy Tyrkalo
Department of Business and Environmental Expertise of Goods
Lviv Polytechnic National University
Lviv, Ukraine
0000-0003-2535-4238

Anatolii Pavlenchyk
Department of Informatics and Kinesiology
Ivan Boberskyi Lviv State University of Physical Culture
Lviv, Ukraine
0000-0002-2205-1883

Volodymyr Dalyk
Department of Management and International Business
Lviv Polytechnic National University
Lviv, Ukraine
0000-0003-0004-2270

***Abstract* — This paper is the development of a crack pipeline defect detection software from the inside for real-time damage assessment using trained model. Collected data on the number of pixels found in the frame that signals damage using a ready-made camera from the JetRacer Kit. The analysis and display of the results of the detected cracks are performed using visualization by a rectangle around damaged concrete. The possibilities of evaluating detected damage based on data on pixels filling a crack in the pipeline are described. The quality of detection of cracks in pipes with artificial and natural lighting is checked.**

Keywords* — *computer vision, machine learning, crack, pipeline, Nvidia Jetson Nano

I. Introduction

Water and sewage pipelines ensure the vital activity of cities and villages. Over time, the water supply system and concrete structures in the form of a pipe wear out, which causes damage during constant use and exposure to external influences. Damage to pipelines from the outside is easy enough to assess, while from the inside it is considered more difficult to access the place, given the dark and size. Due to the small diameter of some places of such pipes, access to determine the visual location of damage in them is limited.

The goal of the work is to implement a program for detecting cracks and further assessing the criticality of the condition of the inner part of the pipeline using a trained model. The main tasks in the work are the implementation and determination of the features of the trained model for the automated detection of cracks in the internal surfaces of pipelines using the JetRacer Kit.

II. Methods and Tools

A. Hardware

The hardware part of the work is based on the ready-to-use device JetRacer Pro 2GB AI Kit [1]. The main element of this Robotic Operating System (ROS) device is built on the basis of NVIDIA Jetson Nano 2GB. The installed operating system is Linux in graphical mode. This allows you to connect to the monitor saved files taken from the video camera through video players and execute commands through the appropriate terminal.

JetRacer Pro kit includes an IMX219-160 camera, the video used to analyze the surfaces of concrete pipelines. The camera has a 160–degree Angle of View (diagonal) and 8 Megapixels. In addition, it contains Brushed motor Electronic Speed Controller (ESC), Carbon Brushed motor and double differentials to overcome minor obstacles. Power is provided by 4 18650 batteries. Control can be performed using a joystick. In addition, a small, attached flashlight is used for lighting inside the pipeline.

B. Software Architecture

Conventionally, the software implementation can be divided into several layers, such as: model training, data recording, visualization of results.

The main part of the software is a deep learning model, which provides the ability to recognize cracks in specific structures in real time. The network architecture is built according to the U-shaped encoder-decoder network architecture (UNET), which consists of encoder and decoder blocks connected by a bridge. The model encoder contains twice the number of filters at each level with two 3x3 convolutions and a Rectified Linear Unit (ReLU) activation function. At each stage, we classify each pixel in more detail to highlight cracks, reducing the size of the image and increasing its depth. The bridge between encoder and decoder contains two 3x3 convolutions and a ReLU activation function. The decoder includes starting from a 2x2 transposed convolution to a 1x1 convolution and sigmoid activation to segment the crack in the input image [2, 3]. Selected model for training on as one of the popular ones for detection based on pattern pictures [3]. The model was trained with an accuracy of 87%.

Visualization of the results in the form of highlighting pixels with red dots and grouping into a square the zone of the highest concentration is performed using the OpenCV library.

The Python programming language was used to develop the program, taking into account the possibility of integration with libraries for working with neural networks. This

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

programming language is installed on the Jetson Nano with the following additional libraries:

- Pytouch, sklearn – deep learning and machine learning libraries
- Onnxruntime, onnxoptimizer, OnnxTransformer – inference engines for Open Neural Network Exchange (ONNX) models
- skl2onnx – converts scikit-learn models to ONNX

III. Implementation

Among the existing developments are implementations of detection of cracks in concrete by various methods [4, 5]. Compared with the review of the latest methods of crack detection by artificial intelligence methods [6], this topic and work is relevant and original, and the scope of application, the specificity of crack environments, and detection methods are different. Pipes can be considered a more difficult object for crack detection due to their difficult accessibility and lack of natural lighting.

All collection and analysis operations take place at Jetson Nano. The main object of research is defects in the inner part of pipelines (as shown in Fig.1), which can be difficult to access and cause the destruction of part of the pipeline.

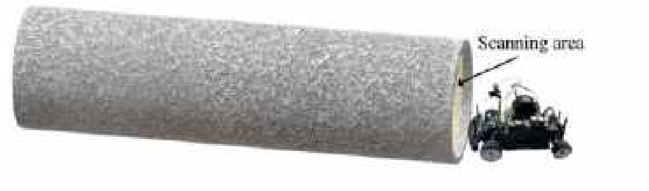


Fig. 1. Example of visualization of the work of the system for finding defects in pipelines using JetRacer Kit

Detection of surface defects in the middle of the pipeline can be performed in the unfilled cavity of the pipe by the method of video analysis while the JetRacer Kit is moving at a low speed.

A. Data collection

The model is trained on the basis of training sets of images with cracks. In addition to datasets from open-source datasets, a dataset from some of our own images of cracks in concrete has been expanded [7]. A feature of pipelines in comparison with a flat surface (road, wall) [8] is their often-round shape. Also, to ensure better accuracy of the model, random rotations and translations are used in the set of images. The neural network weights were saved after training using the PyTorch library. The model was trained on PyTorch and the weights were preserved. The next step is to reduce the model to Open Neural Network Exchange (ONNX) format, an open standard for converting machine learning models into one format [4].

B. Grouping of damage section

Since the image is analyzed for cracks in motion, each frame contains a different number of found damaged areas in the pipeline. After recognition, as shown in Fig. 2, we get the found cracks highlighted by dots. To assess the criticality and damage area, grouping can be performed to visually highlight the pipeline damage area.

To highlight certain objects, such as a face, a car, or others, a rectangle shape is used for recognition. In the case of work with the search for defects in materials, a similar approach can be used. The rectangle surrounding the damaged area shows the mastabas and identified critical sections of the pipeline with the signature of the central point in x and y coordinates.

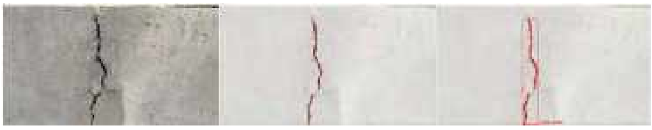


Fig. 2. Example of evaluation crack detection with trained model using test image

Since the points in the coordinates are stored in a different order, you need to set the points around which the rectangle will be drawn. The search for points is carried out according to the following algorithm: sorting the set of points and searching for the minimum and maximum value along the x and y axes along which a rectangle is selected.

Algorithm of crack segmantation using JetRacer Kit

In general, the process of detecting cracks in the pipeline in the process of processing the video stream based on the trained model is shown in Fig. 3.

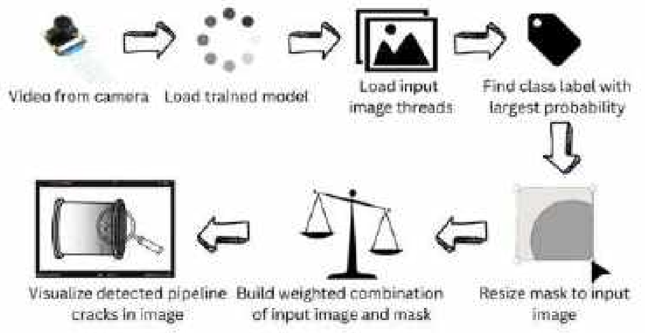

Video from camera
Load trained model
Load input image threads
Find class label with largest probability
Resize mask to input image
Build weighted combination of input image and mask
Visualize detected pipeline cracks in image

Fig. 3. Algorithm of processing data using JetRacer Kit

The program starts execution automatically when the device is started. Each step in the cycle is executed until the application stops and is saved on the device.

As shown in Fig. 3, the video processing algorithm consists of the following steps:

1) Loading model for cracking detection which is trained and converted to ONNX format.
2) Loading the input image threads with resizing it to fit our model parameters.
3) Running forward pass the model.
4) Finding the largest probability in class label for every pixel in the image
5) Resizing the mask to match its dimensions with the input image
6) Merging weighted combination of input image along with mask to form output visualization.
7) Video recording of selected cracks with a damaged area in the pipeline

IV. Experiments

Checking the operation of the proposed program in pipelines can be performed in places with and without natural light. The selection of materials for research consists in finding pipes in different conditions and defects. Most pipelines that are not filled with liquid may contain debris and be in poor technical condition. It is proposed to check the operation of the program and establish the features of defect detection in the dark and during the day.

A. Detection crack in concrete pipelines in the dark

The results of work in the dark can be similar to testing a closed and long pipeline due to the artificial lighting used (a flashlight attached to the device). There may be debris and cobwebs in the pipeline, which makes it difficult for the robot to move. However, these obstacles are not recognized as cracks in the pipeline.A feature of recognition at night is the presence of shadows from objects in the pipe, which affects point incorrect detections. As shown in Fig. 4, a small crack from the shadow of the web was found in the lower part of the pipe

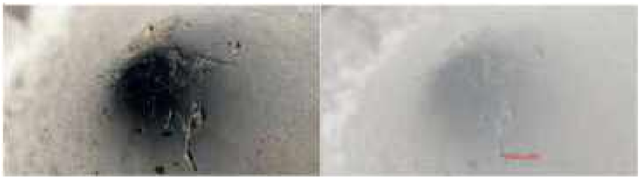


Fig. 4. Input and processed image without cracks in the dark time of day

Another case in Fig. 5 shows the detection of the edge from a crack in a pipeline with a plant.

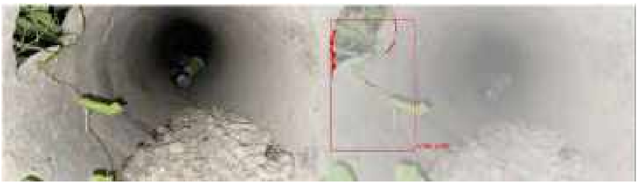


Fig. 5. Input and processed image with detected cracks in the dark time of day

The plant was not recognized, but the shadow from the precise stems formed by the artificial lighting was added to the identified defect section.

B. Detection cracks in concrete pipelines with natural lighting

When scanning the pipeline in daylight, there are no shadows from other objects. Cracks are detected in motion, but the accuracy of identifying a complete defect is different. When approaching the object, the neural network is more clearly able to highlight damage in the pipeline. As shown in Fig. 6, when the robot moves, the plant does not create a shadow. A small part of the crack was also found.

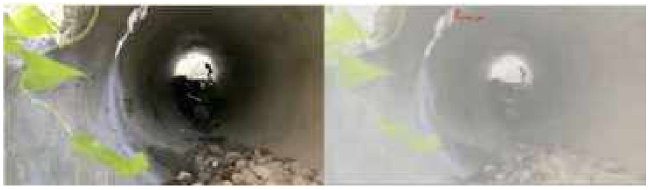


Fig. 6. Input and processed image in daylight is far from damage section

In Fig. 7 shows that when approaching the crack, it more accurately identifies the crack. Also, with the help of the program, the diagnosis of a fairly new pipe was carried out for the presence of cracks.

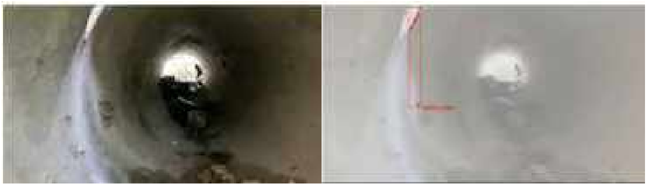


Fig. 7. Input and processed daylight image is closer to the damage section

As shown in Fig. 8 cracks were detected at the beginning of the pipe. The edges of the pipe contained small cracks that were highlighted as a damaged area.

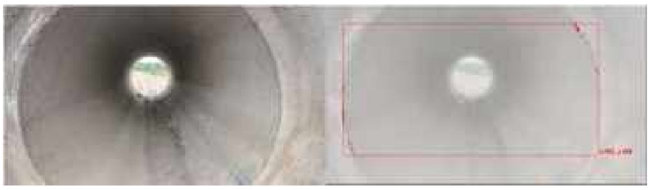


Fig. 8. Input and processed image in daylight with cracks at the junction of pipes.

In addition to the experimental scenarios with cracks, a study was conducted on a section of the pipe without any damage. In Fig. 9 shows a pipe without cracks and without detected defects inside.

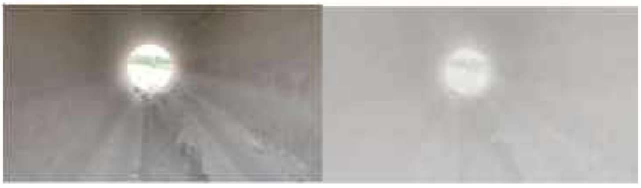


Fig. 9. Input and processed image in daylight without cracks

C. Assessment of damaged sections in pipelines

Based on the number of detected pixels, you can make an analysis and graphically display the dependence of the change in the amount of damage per frame. In this way, it is possible to highlight the main places with cracks along the length of the pipeline. In Fig. 9 shows a pipe without cracks and without detected defects inside

TABLE I. PERCENTAGE OF PIPELINE DAMAGE CRITICALITY PER VIDEO FRAME

Type of lighting	Results with found cracks		
	Video frame	*Percentage of pixels, %*	*Found pixels*
Artificial	Picture 4	0.004	46
	Picture 5	0.209	2462
Natural	Picture 6	0.011	232
	Picture 7	0.031	431
	Picture 8	0.039	516
	Picture 9	0	0

In addition to the experimental scenarios with cracks, a study was conducted on a section of the pipe without any damage. Such results can be useful for analyzing the condition of the pipes as a whole and for assessing the criticality of repairs. The impact of lighting in such enclosed spaces affects the quality of damage detection. In the dark, in the presence of obstacles, debris or cobwebs, additional defects in the pipe may be detected.

Here, in addition to the implementation of the application, it is necessary to take into account methods of monitoring the current state and parameters during data processing [9, 10], quality criteria of approaches [11–17] and research of operations in related possible areas of application [18, 19]. It is also worth considering investment decision support tools [20, 21], management of business structures [22, 23] to ensure efficiency and effectiveness in conditions of risk and complexity of using information technologies and systems.

CONCLUSIONS

Using the JetRacer Kit, a program was implemented to detect cracks inside pipelines using machine learning. In this way, it is possible to carry out diagnostics to hard-to-reach places and assess damage to the pipe surface from the inside.

Crack segmentation is performed in real-time using motion capture from a robot camera on a trained model. Grouping and selection of damage concentration areas can be used to localize damage locations.

The operation of the program was tested and the features of pipe damage detection in the light and dark time of the day were highlighted. The method of assessing the criticality of damage to pipeline sections based on data on pixels filled with cracks is described.

As one of the possible applications, it can be automatic filling with liquid to strengthen the found cracks for longer operation, expanding the capabilities of the system. The obtained results can be useful for the analysis of pipelines for enterprises that are engaged in the maintenance of underground communication networks of cities or villages.

REFERENCES

[1] JetRacer Pro 2GB AI Kit, High Speed AI Racing Robot Powered by Jetson Nano 2GB, Pro Version. [Online]. Available: https://www.waveshare.com/jetracer-pro-2gb-ai-kit.htm

[2] U-NET Architecture Explained and Implementation. [Online]. Available: https://becominghuman.ai/u-net-architecture-explained-and-implementation-470a5095ad57

[3] M. Naqvi, P. P. Sujith, S. Naidu, K. Thomas and U. Ananthanagu, "Exploring Artificial Neural Networks in Virtual Reality: A Unity and Leap Motion-Based Visualization of ONNX Models," 2023 9th International Conference on Virtual Reality (ICVR), Xianyang, China, 2023, pp. 200-204, DOI: 10.1109/ICVR57957.2023.10169803.

[4] V. P. Golding, Z. Gharineiat, H. S. Munawar, and F. Ullah, “Crack Detection in Concrete Structures Using Deep Learning”, Sustainability, vol. 14, no. 13, p. 8117, Jul. 2022, DOI: 10.3390/su14138117

[5] K. Liu, X. Han, and B. M. Chen, “Deep Learning Based Automatic Crack Detection and Segmentation for Unmanned Aerial Vehicle Inspections”, 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO), Dec. 2019, DOI: 10.1109/robio49542.2019.8961534

[6] Hamishebahar Y, Guan H, So S, Jo J. , “A Comprehensive Review of Deep Learning-Based Crack Detection Approaches. Applied Sciences”, 2022, vol. 12, no. 3, p.1374. DOI: 10.3390/app12031374

[7] Surface Crack Detection . [Online]. Available: https://www.kaggle.com/datasets/arunrk7/surface-crack-detection

[8] R. Mysiuk et al., “Video-based Concrete Road Damage Assessment Using JetRacer Kit”, 2023 17th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), Jaroslaw, Poland, 2023, pp. 1–4, DOI: 10.1109/CADSM58174.2023.10076528

[9] R. V. Mysiuk et al., “Determination of conditions for loss of bearing capacity of underground ammonia pipelines based on the monitoring data and flexible search algorithms”, Archives of Materials Science and Engineering, vol. 115, no. 1, pp. 13–20, May 2022, DOI: 10.5604/01.3001.0016.0671

[10] R. Mysiuk, V. Yuzevych, B. Koman, and M. Yasinskyi, “High Availability System for Monitoring Material Degradation Processes at the Concrete-polymer Interface”, 2022 12th International Conference on Advanced Computer Information Technologies (ACIT), Sep. 2022, DOI: 10.1109/acit54803.2022.9913086

[11] L. Yuzevych, R. Skrynkovskyy, and B. Koman, “Development of information support of quality management of underground pipelines”, EUREKA: Physics and Engineering, vol. 4, pp. 49–60, Jul. 2017, DOI: 10.21303/2461-4262.2017.00392

[12] V. Lozovan et al., “Forming the toolset for development of a system to control quality of operation of underground pipelines by oil and gas enterprises with the use of neural networks”, Eastern-European Journal of Enterprise Technologies, vol. 2, no. 5 (98), pp. 41–48, Apr. 2019, DOI: 10.15587/1729-4061.2019.161484

[13] V. Yuzevych, R. Skrynkovskyy, and B. Koman, “Intelligent Analysis of Data Systems for Defects in Underground Gas Pipeline”, 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Aug. 2018, DOI: 10.1109/dsmp.2018.8478560

[14] R. Dzhala et al., “Simulation of Corrosion Fracture of Nano-Concrete at the Interface with Reinforcement Taking into Account Temperature Change”, 4th International Workshop on Modern Machine Learning Technologies and Data Science, MoMLeT&DS 2022, CEUR Workshop Proceedings 3312, Leiden–Lviv, The Netherlands–Ukraine, pp. 123–133, Nov., 25–26, 2022. [Online]. Available: https://ceurws.org/Vol3312/paper10.pdf

[15] L. Yuzevych et al., “Improvement of the toolset for diagnosing underground pipelines of oil and gas enterprises considering changes in internal working pressure”, Eastern-European Journal of Enterprise Technologies, vol. 6, no. 5 (102), pp. 23–29, Nov. 2019, DOI: 10.15587/1729-4061.2019.184247

[16] L. Yuzevych et al., “Improving the diagnostics of underground pipelines at oilandgas enterprises based on determining hydrogen exponent (PH) of the soil media applying neural networks”, Eastern-European Journal of Enterprise Technologies, vol. 4, no. 5 (100), pp. 56–64, Jul. 2019, DOI: 10.15587/1729-4061.2019.174488.

[17] V. Lozovan, R. Dzhala, R. Skrynkovskyy, and V. Yuzevych, “Detection of specific features in the functioning of a system for the anti-corrosion protection of underground pipelines at oil and gas enterprises using neural networks”, Eastern-European Journal of Enterprise Technologies, vol. 1, no. 5 (97), pp. 20–27, Jan. 2019, DOI: 10.15587/1729-4061.2019.154999

[18] A. Sumets et al., “Methodological toolkit for assessing the level of stability of agricultural enterprises”, Agricultural and Resource Economics: International Scientific E-Journal, vol. 8, no. 1, pp. 235–255, Mar. 2022, DOI: 10.51599/are.2022.08.01.12

[19] M. Babych et al., “Substantiation of economic efficiency of using a solar dryer under conditions of personal peasant farms”, Eastern-European Journal of Enterprise Technologies, vol. 6, no. 8 (84), pp. 41–47, Dec. 2016, DOI: 10.15587/1729-4061.2016.83756.

[20] R. M. Skrynkovskyi, “Methodical approaches to economic estimation of investment attractiveness of machine-building enterprises for portfolio investors”, Actual Problems of Economics, vol. 118(4), pp. 177–186, 2011.

[21] R. Skrynkovskyi, “Investment attractiveness evaluation technique for machine-building enterprises”, Actual Problems of Economics, no. 7(85), pp. 228–240, 2008

[22] N. Popova et al., “Marketing Aspects of Innovative Development of Business Organizations in the Sphere of Production, Trade, Transport, and Logistics in VUCA Conditions”, Studies of Applied Economics, vol. 38, no. 4, Feb. 2021, DOI: 10.25115/eea.v38i4.3962

[23] N. Popova, A. Kataiev, R. Skrynkovskyy, and A. Nevertii, “Development of trust marketing in the digital society”, Economic Annals-XXI, vol. 176, no. 3–4, pp. 13–25, Aug. 2019, DOI: 10.21003/ea.v176-02

Application of Neural Networks for Processing Information in Redundant Rate Gyroscopes

Olha Sushchenko
Aerospace Control Systems Dept.
National Aviation University
Kyiv, Ukraine
sushoa@ukr.net

Yurii Bezkorovainyi
Aerospace Control Systems Dept.
National Aviation University
Kyiv, Ukraine
yurii.bezkor@gmail.com

Volodymyr Golitsyn
Aerospace Control Systems Dept.
National Aviation University
Kyiv, Ukraine
vova.gol@ukr.net

***Abstract* — This paper deals with techniques for processing information in redundant rate gyroscopes based on neural networks. The algorithm of processing information based on a neural network with back-propagation is represented. The technique of searching failures in redundant rate gyroscopes based on the neural network is represented. Structures of proposed neural networks are grounded. The choice of activation functions is grounded. Results of functioning developed techniques are shown. The efficiency of the proposed technique is grounded. The obtained results can be applied to redundant measuring instruments assigned for operation on unmanned aerial vehicles.**

Keywords* — *rate gyroscope, neural network, processing information, search of failures, improving accuracy

I. Introduction

For a long time, the basis of intelligent data analysis was ordinary mathematical statistics, which is usually useful in the conditions of testing already-formed hypotheses [1]. At the beginning of its development, neural networks in data analysis caused mixed reviews due to their shortcomings, such as the complexity of the structure, too long learning period, and poor interpretability. But they were compensated by the complex positive qualities, such as a low error rate, constant improvement and optimization of various learning algorithms networks, the algorithm for obtaining rules, and algorithm for simplifying networks. This makes neural networks extremely promising directly in the field of data analysis [2], [3].

Computer technologies for automatic intelligent data analysis are booming. This is due to the influx of new ideas coming from the computer field sciences that were formed at the intersection of artificial intelligence, statistics, and database theory. Elements of automatic data processing and analysis become an integral part of the concept of electronic data storage and are often referred to as data mining (extraction of knowledge from data) [3].

Models of neural networks can be conditionally divided into three types such as:

1) direct distribution networks – one of the most common architectures used in the prediction and recognition of images;

2) networks with feedback, which are used to optimize calculations and associative memory;

3) self-organizing networks containing models of adaptive resonance theory and Kohonen models are used for cluster analysis. [3]

Inertial meters consisting of triaxial rate gyroscopes and accelerometers are the most common in industrial navigation applications today. This gives relevance to the study of excessive inertial meters based on the non-orthogonal arrangement of the measuring axes of the three-axis rather than uniaxial or biaxial gyroscopes. According to the above-mentioned concept, it is reasonable to accept a three-cornered pyramid (tetrahedron) or a four-cornered pyramid (octahedron) as the basic geometric figure of an excessive inertial meter. The use of fragments of such shapes is the most expedient, taking into account constructive requirements on mass and dimensions [4], [5].

Improvement of information processing procedures requires using new optimization techniques including neural networks, fuzzy logic, evolutionary programming, and genetic algorithms [6]. Such an approach can be supplemented by the creation of mathematical descriptions of measuring instruments taking into consideration non-linearities of real devices.

Solution of problems dealing with technical diagnostics, control of complex objects, and processing information in measuring systems requires the development and application of intelligence systems. Artificial neural networks can be used for the creation of the above-mentioned systems.

For problems of information processing, neural networks of both feed-forward and back-forward distribution of signals can be used. Preference must be given to neural networks with feedback [7].

As a rule, the neural network consists of neurons with input vectors. Every input vector is scaled in a definite way using weighting coefficients. The set of weighting coefficients of the neuron simulates its memory. Neurons can be considered as processors, which implement calculations of activation functions based on input signals and weighting coefficients [8].

The basic aim of the research is to develop neural networks, which ensure the minimal output error of the measuring instrument.

II. Processing Data by Neural Network

The researched object is an excessive nonorthogonal meter consisting of rate gyroscopes. In this measuring instrument, a quadrilateral pyramid is used as a reference surface. It contains also 5 triaxial inertial measuring blocks or 15 monoaxial sensors, correspondingly. Hence, to calculate angular velocity projections of a moving vehicle, it is essential to implement transformations based on 15×15 matrix. Navigation information accuracy can be improved by algorithmic methods including specific approaches to processing information [9].

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

The usage of the method of data handling technique grounded on neural networks allows us to determine three monoaxial rate gyroscopes, which have the highest accuracy performances, with the aim of using them for further navigational calculations. This approach allows us to avoid some transformations connected with directional cosine matrices. Projections of the angle velocity of a moving vehicle in the navigation coordinate system can be calculated using measurements of selected inertial sensors. The application of this technique provides reducing errors of measurements and decreases the time of processing data [9] – [11].

From the beginning, we will consider information processing by a uniaxial rate gyroscope. Then, the mathematical linear description of the measurement instrument may be presented in the expression

$$\hat{x} = kx + \tilde{x}_0 , \qquad (1)$$

here $\hat{x}$ is the measurement; x is the obtained quantity; k is the size factor; $\tilde{x}_0$ is the zero bias.

The zero bias belongs to the most important error of a rate gyroscope manufactured by MEMS technology. It is common knowledge that the zero bias contains two constituents [9]

$$\tilde{x}_0 = x_0 + x^t_0 , \qquad (2)$$

where x_0 is the displacement caused by deviations from the technical documentation during production; x^t_0 is the displacement caused by the influence of temperature during functioning of the rate gyroscope.

Applying expression (2), relationship (1) can be expressed as the formula

$$\hat{x} = kx + x_0 + x^t_0 , \qquad (3)$$

here $x^t_0 = k_t \Delta t$, where k_t is the conversion factor; Δt is temperature deflection from the characteristic obtained under standard conditions.

The formalized description of the measurement instrument based on relations (1) – (3) is as follows:

$$\hat{x} = kx + x_0 + k_t \Delta t . \qquad (4)$$

Model (4) can be represented as an equivalent neural network. The diagram of this network is shown in Fig. 1.

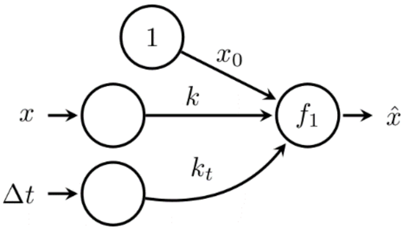


Fig. 1. The neural network for a uniaxial rate gyroscope: f_1 is the activation function.

When a non-linear relationship between the measurement result and the real signal takes place, it is beneficial to utilize approximation based on splines. This allows for the mathematical description of the measuring instrument to be transformed into the following form after some calculations.

$$\hat{x} = k_3 x^3 + k_2 x^2 + k_1 x + k_t \Delta t + x_0 , \qquad (5)$$

where k_1, k_2, k_3 are spline factors.

The neural network that is appropriate to the relationship (5) is shown in Fig. 2 [9].

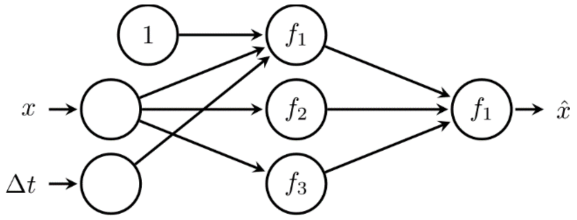


Fig. 2. The neural network for the non-linear model of the object: f_1, f_2, f_3 are single, square, and cubic activation functions.

We can observe in Fig. 2 a neural network, which includes input, hidden, and output layers [12], [13]. These layers utilize activation functions that are differentiated, continuous, and monotonically non-decreasing. To solve the given problem, the backpropagation learning algorithm [13] can be employed, which is commonly used in practical measurement techniques.

The training process of the neural network involves transferring training data to the input layer, propagating the error backward through the network, and adjusting the weight coefficients [14]. The algorithm continues until the root mean square deviation error reaches its minimum in the output layer. The backpropagation algorithm is grounded on the gradient search technique [15].

The output information of non-orthogonal meters intended for work on a moving object (unmanned aerial vehicle) can be presented as a set of equations [9]

$$\begin{aligned} \hat{\omega}_x &= k_x \omega_x + \omega_{x0} + k_{tx} \Delta t; \\ \hat{\omega}_y &= k_y \omega_y + \omega_{y0} + k_{ty} \Delta t; \\ \hat{\omega}_z &= k_z \omega_z + \omega_{z0} + k_{tz} \Delta t, \end{aligned} \qquad (6)$$

where x, y, z are the indices that define angle velocity projections on the axes of the Oxyz system of coordinates.

The diagram of the neural network corresponding to the mathematical description of the measurement block (6) is shown in Fig. 3 [9].

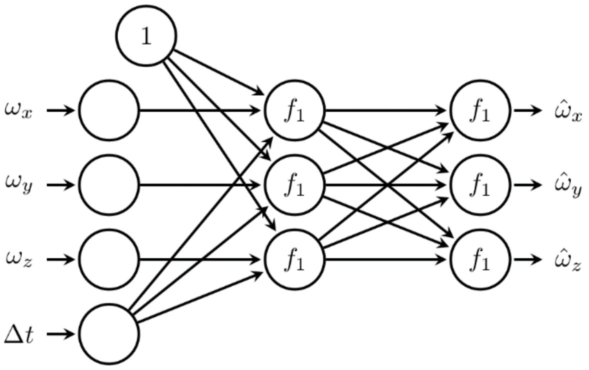


Fig. 3. The neural network of the triaxial rate gyroscope.

The illustration in Fig. 3 is a neural network composed of 3 distinct layers: input, hidden, and output. The input layer is responsible for receiving temperature data and measurements from the measuring instrument. Moving on, the hidden layer processes data from individual three-axis inertial measurement blocks.

Finally, the output layer finds projections of the obtained angle velocity of the mobile vehicles onto the axes of the orthogonal normal system of coordinates.

The neural network depicted in Fig. 4 plays a crucial role in efficiently processing information from an excessive non-orthogonal meter using MEMS sensors. This type of meter presents unique challenges due to its non-orthogonal nature [16].

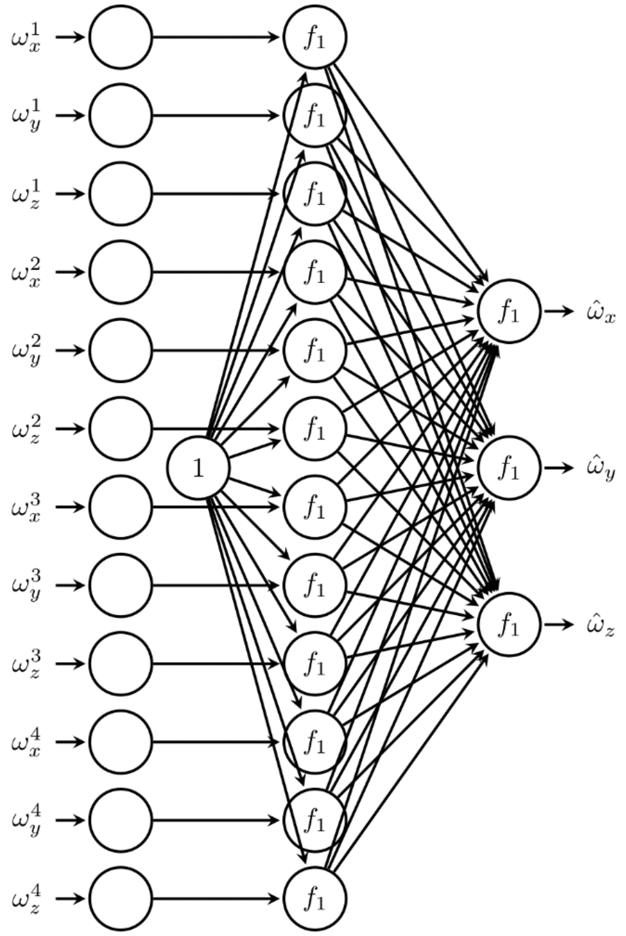


Fig. 4. The neural network for information processing of redundant nonorthogonal meters based on rate gyroscopes.

The network consists of multiple layers that work together to handle the complex data obtained from the MEMS sensors. The input layer receives the raw data from the sensors, which may include measurements from various axes and orientations. This information is then passed on to the subsequent layers for further processing.

The hidden layers in the neural network are responsible for extracting relevant features and patterns from the input data [17] – [19]. They employ sophisticated algorithms and mathematical operations to transform the sensor readings into a more meaningful representation. This step is crucial in capturing the intricate relationships and dependencies within the data [9].

Finally, the output layer produces the desired output based on the processed information. In the case of an excessive non-orthogonal meter, the network's output layer might generate relevant measurements or predictions, such as angle velocity, displacement, or any other relevant parameters.

By leveraging the power of neural networks, this architecture enables accurate and reliable processing of data from an excessive non-orthogonal meter. It effectively addresses the challenges posed by the non-orthogonal nature of the sensor readings, allowing for precise analysis and interpretation of the measurements.

In most cases, rate gyroscopes integrated into inertial measurement units have a built-in thermal stabilization system that minimizes temperature drift. Thus, the temperature-dependent connections within the neural network can be disregarded.

Figure 5 shows an overview of the learning process of a neural network used for handling data from triaxial rate gyroscopes, specifically for *z*-axis of the normal system of coordinates in an excessive non-orthogonal meter [9].

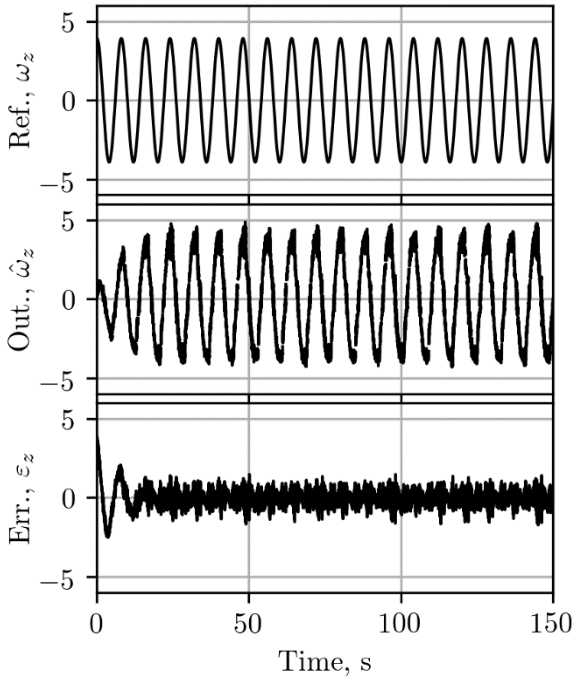


Fig. 5. The learning process of a neural network in for obtaining angle velocity by *z*-axis.

We apply in Fig. 5 such notations as: "Ref" represents the output reference signal of the neural network, "Out" denotes the signal that changes in the learning process, and "Err" indicates the output error.

Figure 6 presents a graph of the evolution of weight factors in the neural network during the training process.

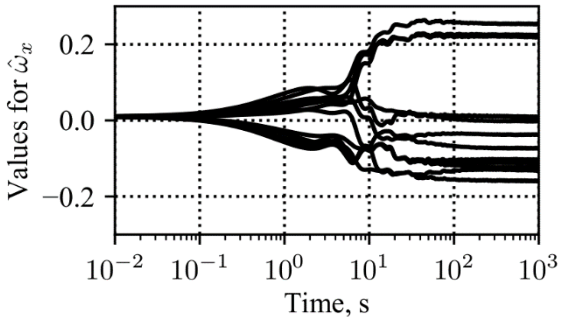


Fig. 6. The variation of coefficients of the neural network for calculating angle velocity projections on the *x*-axis of the normal system of coordinates.

By implementing the proposed information processing algorithm, the selection of optimal three-axis rate sensors is ensured to achieve minimal root mean square errors.

The developed method based on a neural network for information processing allows for the precise determination of the angle velocity projections of a mobile vehicle on the axes of the normal system of coordinates.

The root mean square deflections of the selected rate gyroscopes along the *x*, *y*, *z* axes are characterized by magnitudes of 0.358 deg/s, 0.305 deg/s, and 0.422 deg/s. It should be noted that the experiment was performed under conditions of vibrations of the test bench. The measurement results were subjected to further filtering.

The coefficients obtained during neural network training are shown in Table I.

TABLE I. COEFFICIENTS OF LEARNING NEURAL NETWORK

Input axis ω_j	Output Axes		
	x	*y*	*z*
x_1	-0.234	-0.003	-0.004
y_1	0.021	-0.233	0.023
z_1	0.010	-0.001	-0.234
x_2	0.054	-0.093	-0.192
y_2	0.226	0.090	0.019
z_2	-0.060	0.195	-0.119
x_3	-0.182	-0.103	0.065
y_3	-0.092	0.089	-0.193
z_3	-0.064	0.190	0.115
x_4	0,171	-0.106	0.165
y_4	-0.108	0.089	0.202
z_4	0.148	0.206	0.016
bias	-0.003	-0.005	-0.009

The proposed data handling method significantly improves determining the angle velocity of a mobile vehicle. Unlike the standard information processing procedure, which involves complex calculations that consider the matrix of conversions between measuring and normal systems of coordinates, the proposed method simplifies the calculating process. For most purposes, coefficients of the measurement equation can be determined using linear dependencies (3.6) and the least squares method. However, implementing the technique of least squares is connected with substantial computational burdens, for example, large RAM capacity for accumulation of initial information and intermediate processed data, as well as longer execution time. In contrast, neural networks are not characterized by these limitations and may be realized on microcontrollers with bounded computational burdens, making them a more efficient option for information processing [20], [21].

III. SEARCH OF FAILURES USING NEURAL NETWORKS

The process of developing a fault determination and location method is intricate and involves numerous transformations and calculations [22].

An excessive non-orthogonal meter can be viewed as a collection of individual devices. In this type of neural network, there are 2 layers. Neurons in the 1st layer are connected to the sensors that form part of the excessive non-orthogonal meter.

Let us assume that we have *n* measuring observations. The deflection of the *j*th observation from the average magnitude may be determined using the following relationship:

$$\Delta_j = x_j - \tfrac{1}{n-1}\sum_{\substack{i=1\\ i\neq j}}^{n} x_i \ . \tag{7}$$

To estimate the deviation (7) from the average measurement result, some limiting value ε is used. The condition of sensor performance can be presented in the form

$$\Delta_j^2 \le \varepsilon^2 \ . \tag{8}$$

Expression (8) for estimating the measurement error was chosen taking into account the linear-quadratic approach [23] – [25].

The neural network activation functions may be written in the following way

$$f_1 = x_j^2 \ ; \tag{9}$$

$$f_2 = \frac{1}{1+e^{-y_j}} \ . \tag{10}$$

Expressions (9), and (10) describe the quadratic and sigma functions, correspondingly [14], [15]. The weighting coefficients at the input and output of the 1st layer are supposed to be equal to 1. The weighting coefficients at the output of the 2nd layer are calculated in accordance with the sigma function

$$w_{ij} = \begin{cases} 1, i = j; \\ \dfrac{-1}{n-1}, i \neq j, \end{cases} \tag{11}$$

where $y = \Delta_j^2$.

A neural network (9) – (11) is utilized in this scenario to handle an excessive nonorthogonal configuration resulting from combining a set of *n* inertial sensors. The specific focus is on a measuring device assigned for the measurement of the angle velocity of a mobile vehicle.

For a clearer understanding, the chart of this neural network is depicted in Fig. 7 [22]. This diagram visually represents the components and their connections within the network.

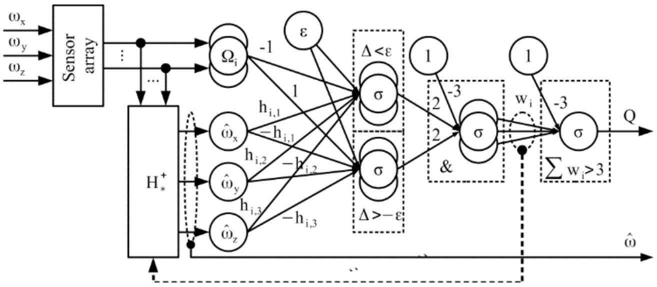


Fig. 7. The learning process of a neural network in for calculation of angle velocity projection on *z*-axis.

Using the matrix of direction cosines, we can calculate projections of the angle speed on the measuring system of coordinates of the non-orthogonal measuring instrument [26]

$$\Omega = H\omega , \tag{12}$$

where Ω is *n*×1 matrix; H is *n*×3 matrix; ω is 3×1 matrix.

In general, the matrix H can be not square. Hence, it is possible to re-establish the magnitude of the reference angle speed using the Moore-Penrose theorem [27] – [29]

$$\hat{\omega} = H_{ps}\Omega_m , \tag{13}$$

where $\hat{\omega}$ is the vector of estimates of the obtained angle velocities; H_{ps} is the pseudo-inverse matrix $H_{ps} = (H^T H)^{-1} H$; Ω_m is the vector of obtained angle velocities.

The error of measurements for each information channel (each rate gyroscope) may be obtained as

$$\Delta = \Omega_i - \Omega . \tag{14}$$

The ground for problem-solving a decision about a malfunction of the information channel is as follows

$$|\Delta| < \varepsilon . \tag{15}$$

The method mentioned above for detecting faults in inertial sensors can be effectively implemented using an equivalent neural network architecture. It is important to highlight that the sigma function is widely used as the most common activation function in this context.

$$\sigma(x) = \frac{1}{1+e^{-kx}} , \tag{16}$$

where *k* is the factor of the gamma function.

To implement the procedure of excluding measurement channels with significant deviations in readings of the rate gyroscope from the measured values of the angle velocity of the moving object, it suits to utilize the equivalent matrix of direction cosines in the following expression

$$H_e = \mathrm{diag}(w_1, w_2, ..., w_n)H . \tag{17}$$

Using formulas (12) – (17) and the Moore-Penrose algorithm [24], [25], it is possible to create a system for estimating the running value of the angle velocity based on information coming only from the "active" axes.

$$\hat{\omega} = H_e(+)\Omega_i . \tag{18}$$

However, for a non-orthogonal measurement system to work properly, three axes of sensitivity at least must be used. Then, the condition of complete failure can be written as

$$\sum w_i > 3 , \tag{19}$$

where w_i is the weight function after reduction to the equivalent layer of the neural network

$$Q = \sigma(\sum w_i - 3) , \tag{20}$$

where *Q* is the probability of failure.

Formulas (18) – (20) ensure estimating angle velocities of the mobile vehicle.

Calibrating and calculating components in the matrix of directional cosines can be achieved by implementing the algorithm proposed in [30]. To train the neural network, the training data is fed into the network inputs, and the error is back-propagated to adjust the elements in the matrix of directional cosines (also known as the weighting coefficients in the neural network). It is important to note that in this process, the obtained angle velocity is considered to be a given parameter.

IV. Case Study

The effectiveness of the proposed approach to fault finding is confirmed by the simulation results presented in Fig. 8.

A neural network can approximate an arbitrary function with a given accuracy. Thus, the AND function can be implemented in the form of a separate layer of a neural network with connection coefficients that allow you to adjust the smoothness of the function. This prevents the occurrence of transient processes when estimating the inertial parameters of motion. This approach also makes it possible to apply a single mathematical apparatus of neural networks with direct communication without the use of a Boolean mathematical apparatus, which simplifies the implementation of this algorithm.

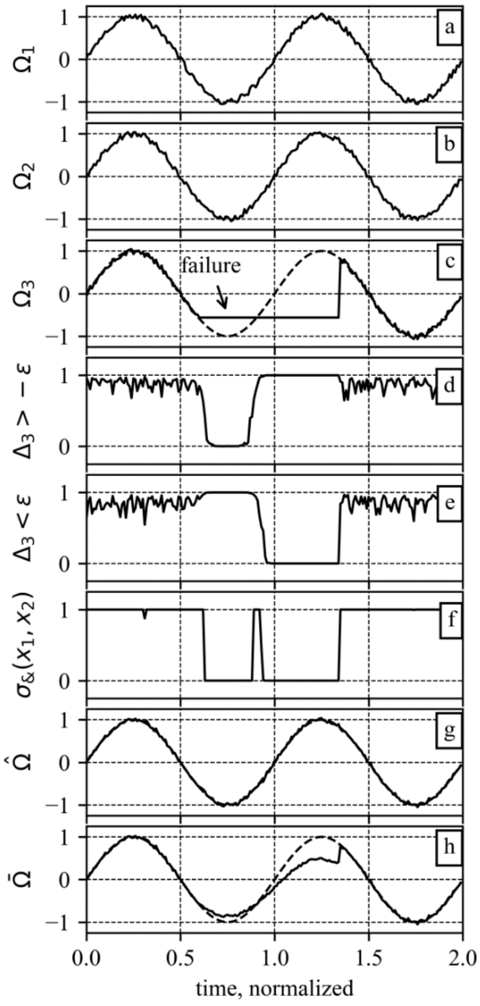


Fig. 8. Modelling results.

Graphs *a*, *b*, and *c* in Fig. 8 depict the simulation of the output signal of a block of three sensors. The graph *c* shows the temporary failure of the sensor in the form of the "sticking" effect of readings. Graphs *d* and *e* illustrate the output of the neural network, which evaluates the deviation of sensor 3 readings beyond the permissible range ε. Graph *f* shows the output from the AND element, which forms a sign of the performance of a specific sensor at the current time (in this case, sensor 3). The graph *g* shows the results of readings formed taking into account the assumption that the readings of sensor 3 are unreliable, and the graph *h* is, accordingly, when this fact is ignored. As can be seen from these graphs, the conventional processing system of the sensor unit with redundancy with the averaging of readings in the presence of failures allows the distortion of the output signal, and the proposed system allows to prevent this event.

Conclusions

This study introduces information processing methods utilizing neural networks to account for temperature influence and sensor parameter deviation. The neural network is trained using samples, reducing calibration time compared to the least-squares algorithm. The learning process of the neural network is demonstrated, yielding conversion weight matrices and normal deflections of selected rate gyroscopes.

Modelling results indicate that applying neural network-based calibration algorithms results in a minimum standard deviation of 0.42 deg/s. This method can be employed in developing medium-accuracy measuring systems, especially for designing UAV measurement systems.

Additionally, an enhanced method based on a neural network is proposed, enabling real-time analysis of measuring channel accuracy and a comprehensive evaluation of the inertial excess meter's performance.

Comparing the classic method of processing navigation information with neural network approaches, it is observed that the calculation time is independent of the measured value type. Consequently, neural network-based processing methods offer optimal computational load, making them highly suitable for UAV applications.

References

[1] R. Xu and D. Wunsch, “Survey of clustering algorithms,” IEEE Transactions on Neural Networks, vol. 16, no. 3, 2005, pp. 645–678.

[2] N. Xianjun, “Research of data mining based on neural networks,” Engineering and Technology, no. 39, 2008, pp. 381-384.

[3] L. Leshchinsky and O. Ishchenko,”Usage of neural networks in process of intellectual (cluster) data analysis,” Mathematical Methods, Models and Information Technologies in Economics,” issue 17, 2017, pp. 578–581. (in Ukrainian).

[4] V. Chikovani, O. Sushchenko, and H. Tsiruk, “Redundant information processing techniques comparison for differential vibratory gyroscope,” Eastern-European Journal of Enterprise Technologies,” vol. 4 (7-82), 2016, pp. 45-52.

[5] O.A. Sushchenko, V.O. Golitsyn, “Data processing system for altitude navigation sensor,” in Proc. IEEE 4th International Conference Methods and Systems of Navigation and Motion Control (MSNMC 2016), Kyiv, Ukraine, 2016, pp. 84-87.

[6] O. Sushchenko, Y. Bezkorovainyi, and N. Novytska, “Theoretical and experimental assessments of accuracy of nonorthogonal MEMS sensor arrays,” Eastern-European Journal of Enterprise Technologies, vol. 3 (9-93), 2018, pp. 40-49.

[7] C.C. Aggarwal, Neural Networks and Deep Learning, Cham: Spinger, 2023, 529 p.

[8] S.O. Subbotin, Neural Networks: Theory and Practice, Zhytomir: O.O.Evenock, 2020, 184 p.

[9] O. A. Sushchenko, Y.M. Bezkorovayniy, and V. O. Golytsin, “Processing of redundant information in airborne electronic systems by means of neural networks,” in Proc. IEEE 39th International Conference on Electronics and Nanotechnology, (ELNANO 2019), Kyiv, Ukraine, 2019, pp. 652-655.

[10] R.H. Rogne, T. H. Bryne, T. I., Fossen, and T.A. Johansen, “Redundant MEMS-based inertial navigation using nonlinear observers,” Journal of Dynamic Systems, Measurement, and Control, vol. 140 (7), 2018, Paper No DS-17-1023, 7 p.

[11] M. Jafari, “Optimal redundant sensor configuration for accuracy increasing in space inertial navigation systems,” Aerospace Science and Technology, 2015, vol. 47, pp. 467–472.

[12] L. Wu, P. Cui, J. Pei, and L. Zhao. Graph Neural Networks: Foundations, Frontiers, and Applications. Springer, Singapore, 2022

[13] B. Mehlig. Machine Learning with Neural Networks. Goteborg, 2021. 260 p.

[14] S. Haykin, Neural Networks and Learning Algorithms. Boston: Pearson, 2008. 936 p.

[15] K.L.Du and N.M.S. Swamy, Neural Networks and Statistical Learning, Berlin: Springer, Science &Business Media, 2013, 824 p.

[16] M. Jafari, “Optimal redundant sensor configuration for accuracy increasing in space inertial navigation systems,” Aerospace Science and Technology, vol. 47, 2015, pp. 467–472.

[17] L.N. Da Silva, D.H. Spatti, K.A. Flaurizino, L.H. Bartocci Liboni, S.F. dos Reis Alvwe, Artificial Neural Networks: a Practical Course, Berlin: Springer 2017, 327 p.

[18] G. Di Franco and M. Santurro, “Machine learning, artificial neural networks and social research,” Quality and Quantity, 2021, vol 55 (6324), p. 1007–1025.

[19] M. Zaliskyi and O. Solomentsev, “Method of Sequential Estimation of Statistical Distribution Parameters in Control Systems Design,” in Proc. IEEE 3rd International Conference Methods and Systems of Navigation and Motion Control (MSNMC), Kyiv, Ukraine, 2014, pp. 135–138.

[20] I.V. Ostroumov, K. Marais, and N.S. Kuzmenko, "Aircraft positioning using multiple distance measurements and spline prediction," Aviation, vol. 26, issue 1, 2022, pp. 1-10.

[21] M. Zaliskyi, Yu. Petrova, M. Asanov and E. Bekirov, “Statistical Data Processing During Wind Generators Operation,” International Journal of Electrical and Electronic Engineering & Telecommunications, vol. 8, no. 1, 2019, pp. 33–38.

[22] O.A. Sushchenko, Y.M. Bezkorovainyi, V.O. Golitsyn, “Fault-tolerant Inertial Measuring Instrument with Neural Network,” in. Proc.IEEE 40th International Conference on Electronics and Nanotechnology (ELNANO), Kyiv, Ukraine, 2020, pp. 797–801.

[23] R.L. Ott and M.T. Longnecker, An Introduction to Statistical Methods and Data Analysis. Boston: Cengage Learning, 2015. 1296 p.

[24] B. Efron and T. Nastie, Computer Age Statistical Inference: Algorithms, Evidence and Data Science. Cambridge University Press, 2016. 495 p.

[25] R.B. Millar, Maximum Likelihood Estimation and Inference: with Examples in R, SAS and ADMB. London: Wiley, 2011. 357 p

[26] X. Dai, L Zhao, and Z. Shi, “Fault tolerant control in redundant inertial navigation system,” Mathematical Problems in Engineering. 2013. Vol. 2013. pp. 1–11.

[27] Rogne R.H. Bryne T.H., Fossen T.I., Johansen T.A. Redundant MEMS-Based Inertial Navigation Using Nonlinear Observers. Journal of Dynamic Systems, Measurement, and Control. 2018. Vol. 140. Issue 7

[28] Y.N. Bezkorovainyi and O.A. Sushchenko, “Improvement of UAV positioning by information of inertial sensors,” in Proc. IEEE 5th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC), Kyiv, Ukraine, pp. 151-155.

[29] O. Solomentsev, M. Zaliskyi, Yu. Nemyrovets, and M. Asanov, “Signal Processing in Case of Radio Equipment Technical State Deterioration,” Signal Processing Symposium 2015 (SPS 2015), Debe, Poland, June 10-12, 2015, pp. 1–5.

[30] O. C. Okoro, M. Zaliskyi, S. Dmytriiev, O. Solomentsev, and O. Sribna, “Optimization of Maintenance Task Interval of Aircraft Systems,” International Journal of Computer Network and Information Security (IJCNIS), 2022, Vol.14, No. 2, pp. 77-89.

Optimization Problems on Complex Networks: Method of Gravitational Potentials

Yuriy Golovaty
Dept. of Mathematical Statistics & Differential Equations
Ivan Franko National University of Lviv
Lviv, Ukraine
yuriy.golovaty@lnu.edu.ua

Oleksa Hryniv
Data Science Office, ELEKS
Ukrainian Catholic University
Lviv, Ukraine
oleksa.hryniv@gmail.com

***Abstract* — We study the problem of building an efficient service network on a complex map of potential customers' locations and develop an iterative algorithm producing nearly optimal service center configurations. We propose a probabilistic model of customers' behavior when they choose service centers. A service network of a given size is optimized according to several criteria that ensure an even distribution of the load on the network, predictability of the number of requests to each center, and convenient location of service centers for customers. This optimization problem is NP-hard, but we propose an efficient greedy algorithm based on the spectral properties of the graph Laplacian. The concept of gravitational potential for a particular service network configuration is introduced. The gravitational potential allows us to calculate customer preferences and metrics that measure network efficiency. In addition, the gravitational potential is the basis of our greedy algorithm, that rebuilds the network at each iteration step to achieve better efficiency. Our research was inspired by a recent work of Steinberger [1] on the graph's shortest path problem.**

Keywords — complex network, graph Laplacian matrix, spectral method, optimization, probabilistic model, customer preference, greedy algorithm, service network.

I. Introduction

Statistical data, which in the last century were large sets of numbers, are now increasingly taking on the form of non-trivial geometric structures. The processes taking place in social, biological, technological, transportation and other infrastructure networks, etc., can be effectively studied only by considering the complex geometry of these structures [2], [3]. Mathematical and statistical models on complex networks are being actively researched in the 21st century because they require the development of new methods and algorithms. Optimization problems on graphs are an exciting and challenging class of new problems, most of which involve NP-hard algorithms [4]. Usually, the term 'network' is associated only with the computer hubs, and thus 'network optimization' is regarded chiefly as a boost of the server's data processing efficiency (see, for instance, [13]). However, complex networks and graph optimization problems are applied to many situations, most of which have nothing to do with computer networks.

Among the various methods for studying complex networks, spectral methods occupy a significant place. Structural matrices of graphs, such as adjacency matrices or Laplace matrices, contain information about the topology of graphs [5]. With an increase in the number of vertices and edges of a graph, this information becomes more and more manifested in the spectra and eigenspaces of these matrices. We can apply spectral methods to various problems dealing with graphs, such as clustering [6], [7], [8], community detection [9], machine learning and pattern recognition, dimensionality reduction and data representation, graph drawing [10], [11], [12].

In this article, we apply spectral analysis techniques to a problem of service network optimization. The task is to find the most efficient location of service centers so that the number of customers' requests would be predictable, the workload on the centers would be uniform, and the centers would be close to customers. We built a probabilistic model of customer behavior when choosing one of the service centers on the location map. Regarding probabilistic distributions of customer preferences, we formulated the criterion of network optimality for a predetermined number of service centers.

Applying variational properties of the graph Laplace operator, we introduce the concept of gravitational potential and propose a method for its construction for each specific location of the service network. Using this potential, we can calculate the network optimality metrics and propose an algorithm to reduce the values of these metrics effectively. Of course, we can find a solution (an optimal configuration of the service network) for every reasonable optimality criterion since we are dealing with minimizing a function on a finite set. This solution is not, in general, unique due to symmetries of the graph. However, the brute-force method leads to an NP-hard problem, and the computation time grows dramatically with the number of graph vertices. We propose a new spectral network optimization method based on the paradigm of greedy algorithms. A greedy strategy does not always lead to an optimal solution. Nevertheless, the computational efficiency of our algorithm and the experimentally established deviation from optimality within 10-15% justify its use. The algorithm performs clustering of the network into groups of customers for each service center and then determines each center's optimal location.

II. Problem Statement

Let $G = G(V, W)$ be a weighted, undirected, connected graph, where $V = \{v_1, \dots, v_N\}$ is a set of vertices, and W is a weight matrix. The vertices of G represent locations where potential customers (clients) are situated. The symmetric matrix $W = \left(w_{ij}\right)_{i,j=1}^{N}$ represents a map of roads between locations with distances between them: if $w_{ij} > 0$, then the locations v_i and v_j are connected by a road of length w_{ij}. We assume that customers are uniformly distributed across these locations. Our goals are to construct an efficient

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

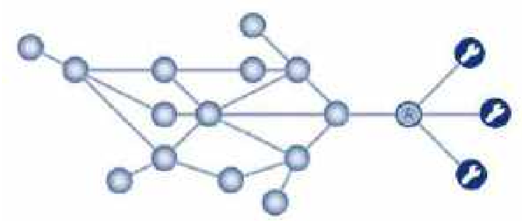


Fig. 1. A service network in which a bottleneck separates customers and service centers.

service network on such a location map if the number of service centers is predetermined. Provided that all service centers have the same performance, we optimize the service network according to the following criteria: the uniform load on all service centers, a large share of regular customers for each service point, which ensures predictable behavior of customers, and the convenient location of the whole network, namely, each service center should be as close as possible to its regular customers.

Assume that the service network consists of K centers, and K is much smaller than $N = |V|$. Let us choose a subset

$$S = \{s_1, \dots, s_K\} \subset V$$

of K vertices. By placing the service centers in the nodes of S, we will get $M = N - K$ locations $v_1, \dots, v_M$ without their service centers. It is worth noting that only these locations affect the optimization criteria. Indeed, the customers from S are uniformly distributed across service centers due to our assumptions, and they are obviously regular customers of the service point at their location. Customer behavior is influenced by random factors, and therefore we are dealing with a probabilistic model. Suppose p_{ij} is the probability that the customer from the location $v_i \in V \setminus S$ chooses the service center s_j; we assume this probability is the same for all customers from the location v_i. Let us combine all probability distributions of the customers' choice in the matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \cdots & \cdots & \cdots & \cdots \\ p_{M1} & p_{M2} & \cdots & p_{MK} \end{pmatrix}.$$

Note that the matrix depends on S. A change in the service network affects the user's preferences. An algorithm for constructing P for each subset S is proposed in Sec. III.B. The algorithm is based on the spectral properties of the discrete Laplace operator on G.

The sum of the entries in each row of P is equal to 1, while the sum of the entries in the j-th column

$$E_j = \sum\nolimits_{i=1}^{M} p_{ij}$$

is the expected loading of the service center s_j. Note that

$$\sum\nolimits_{j=1}^{K} \frac{E_j}{M} = 1,$$

since $\sum_{j=1}^{K} E_j = \sum_{i=1}^{M} \sum_{j=1}^{K} p_{ij} = M$. Therefore, $\frac{E_j}{M}$ is a share of customers that the service center s_j must serve. In the case of the optimal network, this ratio should be as close as possible to $\frac{1}{K}$. Let us introduce the first optimization criterion

$$J_1 = \left(\sum\nolimits_{j=1}^{K} \left(\frac{E_j}{M} - \frac{1}{K}\right)^2\right)^{1/2}.$$

Given $S \subset V$, the value $J_1(S)$ indicates how close the load distribution for service centers is to uniform.

This metric is realistic and intuitive. However, we can achieve uniform load distribution for inefficient networks as well. For example, consider the situation with the "bottleneck" shown in Fig. 1. Customers who want to contact a service center marked with a wrench must go through vertex A. At this vertex, they must decide which service center they will choose. Since we assume that centers are identical, the customers do not prefer any of the centers and choose them with equal probability. Hence $p_{ij} = \frac{1}{K}$ and, consequently, $J_1 = 0$. The law of large numbers ensures that customers should be evenly distributed between service nodes on average (if we observe the network for a long time). But this will not help us on some day when several centers will suddenly have long queues while others will be waiting for customers. This network is inefficient because none of the centers have regular customers and cannot predict the number of requests.

Let us consider the standard simplex

$$\Delta = \{x \in \mathbb{R}^K : x_1 + x_2 + \dots + x_K = 1,\ x_i \geq 0\}$$

in $\mathbb{R}^K$. The probability distribution of each customer is a point in this set. The simplex has K vertices

$$\pi_1 = (1,0,\dots,0),\ \pi_2 = (0,1,0,\dots,0), \dots, \pi_K = (0,\dots,0,1)$$

that correspond to distributions of regular customers. Denote by $P_1, P_2, \dots, P_M$ the rows of P. Let us visualize the choice of our clients by drawing all points P_i on Δ. Fig. 2 shows such a visualization in the case $K = 3$ for two different networks optimal according to the criterion J_1. The first network is like a network with a bottleneck since all probability distributions are located around point $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. This point is in the zone of choice uncertainty. In another network, the clients are divided into three approximately equal parts, and their distributions lie in a vicinity of the vertices π_1, π_2, and π_3. In such a network, the behavior of customers is well predictable.

We will add one more optimality criterion for the optimal networks. Let π be the set of vertices $\pi_1, \dots, \pi_K$. We introduce the metric

$$J_2 = \frac{1}{M} \sum\nolimits_{i=1}^{M} dist(P_i, \pi),$$

where

$$dist(P_i, \pi) = \min_{j=1,\dots,K} \|P_i - \pi_j\|_{\mathbb{R}^K}.$$

Finally, the criterion for the optimality of the service network will be the average value of the two previous ones

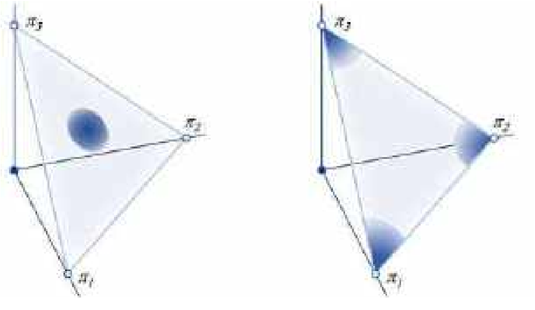


Fig. 2: Visualization of the customer's choice on the probabilistic simplex.

$$J(S) = \sqrt{\frac{K}{K-1}}\ \frac{J_1(S)+J_2(S)}{2}$$

with the normalized factor, due to which the metric J takes values in [0, 1].

Our algorithm applied to the optimization problem

$$S_* = \arg\min_{S\subset V} J(S)$$

finds the optimal (or near-optimal) location S_* of the service network and groups customers into K clusters

$$C_1, C_2, \dots, C_K.$$

Each cluster C_j consists of locations where customers are most likely to choose the service center s_j. Returning to our visualization, C_j consists of those customers whose distributions P_i lie in a neighborhood of the vertex π_j. In the last stage, we minimize the average value of customers' shortest paths to their service centers in each cluster. When we find such a place in the cluster, we will move the service center there. Of course, such a change can affect the metric J, but this effect is insignificant. A more convenient location of the service center in the cluster only increases its attractiveness to customers. Only customers on the edge of clusters can change their choice. However, there are few such customers because of the optimization by J_2.

III. Gravitational Potentials and Customers' Choice

We denote by L the graph Laplacian of G. This operator is a matrix of order n defined by $L = D - W$, where D is the diagonal matrix $\mathrm{diag}(d_1, \dots, d_n)$ with

$$d_i = \sum_{j=1}^{n} w_{ij}.$$

The Laplacian matrix has several interesting properties, which can be applied to graph analysis: L is positive semidefinite, L has always zero eigenvalue, the multiplicity of this eigenvalue is equal to the number of connected components of the graph G, and therefore G is connected if and only if the zero eigenvalue is simple. The spectrum of the Laplacian and its eigenvectors contain information about the topological structure of G

A. Gravitational potentials

The method of gravitational potentials we propose in this article is also a spectral method. Using variational and spectral properties of the graph Laplace operator, we construct a nonnegative function $\Psi_S: V \to \mathbb{R}$ on vertices of G, which we call the *gravitational potential*. The potential Ψ_S is equal to zero only at vertices of S. At the rest of the vertices, the potential is positive, and its value $\Psi_S(v)$ increases with the distance from the vertex v to the set S. It is natural to think of Ψ_S as a potential of the gravitational field, where service centers are sources, which produces a force on other vertices. In Fig. 3, we show a simple and somewhat naive interpretation of this potential. The service centers are global minima of Ψ_S, each attracting a certain number of customers if they prefer to move along gradient flows. Given $S \subset V$, the gravitational potential allows us to calculate the matrix P of customers' choice and the J-metric, which measures network efficiency. In addition, the potential is the basis of our greedy algorithm, as it shows how to rebuild the network at each iteration step to reach the smallest value $J(S)$.

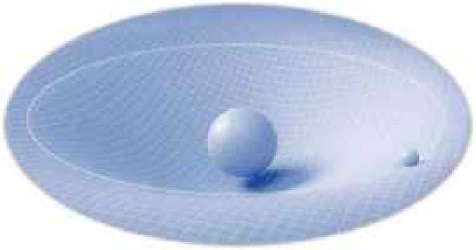


Fig. 3: The service center (giant planet) creates a gravitational field that retain a customer (minor planet).

Let us describe the algorithm for constructing the gravitational potential. We start with the Laplace matrix L. Each row (column) of L corresponds to a particular vertex of G. Let us remove from L all the rows and columns that correspond to the vertices $s_1, \dots, s_K$. We denote the reduced matrix by L_S. It is a square matrix of size $M = N - K$. We next find the eigenvector $\psi \in \mathbb{R}^M$ associated with the smallest positive eigenvalue of L_S. The entries of this eigenvector have the same sign. Thus, we can assume ψ to be non-negative. If all entries of ψ are positive, then we set

$$\Psi_S(v) = \begin{cases} \psi(v), & \text{if } v \in V \setminus S, \\ 0, & \text{if } v \in S. \end{cases}$$

This algorithm works only if the graph G remains connected after removing the vertices $s_1, \dots, s_K$ from it. Otherwise, the eigenvector ψ has zero coordinates.

Suppose the graph G splits into the connected components

$$G_1(V_1, W_1), \dots, G_m(V_m, W_m)$$

after removing S. Then $L_S: \mathbb{R}^M \to \mathbb{R}^M$ is the direct sum

$$L_1 \oplus \cdots \oplus L_m$$

of linear maps $L_k: \mathbb{R}^{M_k} \to \mathbb{R}^{M_k}$, where $M_k = |V_k|$. That is, the matrix L_S can be made block-diagonal by renumbering the vertices, if necessary. Let ψ_k be the positive eigenvector that corresponds to the smallest positive eigenvalue of L_k. We set

$$\Psi_S(v) = \begin{cases} \psi_1(v), & \text{if } v \in V_1, \\ & \dots \\ \psi_m(v), & \text{if } v \in V_m, \\ 0, & \text{if } v \in S. \end{cases}$$

The gravitational potential Ψ_S has the following property.

> *Each vertex $v \in V \setminus S$ is connected to some other vertex $u \in V$ such that $\Psi_S(u) < \Psi_S(v)$.*

B. Construction of the matrix P

Knowing the gravitational potential, we can construct the matrix P describing all customer's probability distributions. These distributions determine the customers' choice of service centers. We find the paths from the customer to the service centers on the graph G to calculate the probabilities. However, we will only consider some of the paths, but only the shortest ones regarding the potential Ψ_S. Building paths can be viewed as a discrete analog of the gradient descent method, where we choose the direction that minimizes Ψ_S the most on each step.

Let's choose a customer and take him to one of the service centers. Since graph G is connected, the set of neighboring vertices for the vertex with our client is not empty. Among these vertices, we will choose the one for which the value of the gravitational potential is the smallest. Go to the selected vertex and repeat the previous step. So, step by step, we will move toward decreasing Ψ_S until we reach the zero level, where only service centers are located. The property of the gravitational potential described above is essential to prove that this algorithm works correctly, and we will find such a path. Indeed, we "lose height" at each step. Since graph G is finite, we will reach a vertex with a zero potential value (i.e., a place with a service center) in a limited number of steps. It is worth noting that the choice of the next vertex on the path is sometimes ambiguous because the gravitational potential can reach a minimum value in two or more neighboring vertices.

Consider a customer who is in location v_i. If there is a unique path for this customer and it leads to the service center s_j, then we set $p_{ij} = 1$. This means that $P_i = \pi_j$, i.e., we are dealing with a regular customer of s_j. Suppose the customer has n different paths to the service centers, one of which is s_j. Then we put

$$p_{ij} = \frac{n_j}{n},$$

where n_j is the number of paths to s_j. If the customer cannot get to s_j via any of these paths, then $p_{ij} = 0$.

IV. Optimization Algorithm

Let us start the search for an efficient service network by randomly selecting a set $S_0 \subset V$ consisting of K vertices of the graph G. We now can construct the gravitational potential Ψ_{S_0} and the matrix $P(S_0)$. Then, having the probability matrix, we can calculate the optimization criterion's value $J(S_0)$. The optimization algorithm based on the greedy strategy is as follows. At each iteration step, we will move one of the service centers to another location to reduce the J-metric. We will again refer to the potential Ψ_{S_0} and find the vertex

$$v_* = \arg\max_{v \in V} \Psi_{S_0}(v),$$

where it acquires the highest value. If there are several such vertices, we choose any of them. Let us move the service centers, which gives the best value of the J-criterion, to the vertex v_*. Now we have a new configuration $S_1 \subset V$ of the service network with $J(S_1) < J(S_0)$ and proceed to the next iteration step. On each iteration we pick a node with the highest value of Ψ_S and move one of the service centers to it. We continue this process, reducing the deviation from the optimum.

Of course, we do not expect this process to converge. There are both mathematical and algorithmic reasons for this. The solution of optimization problem is generally not unique, and greedy algorithms on complex networks rarely give an exact solution. However, their computational efficiency makes them worthwhile even if they provide only approximate results. We continue the iteration process until the optimization score no longer decreases.

Let S_* be the best location for the service network according to our greedy algorithm. By computing the potential Ψ_{S_*} and the customer choice matrix $P(S_*)$ again, we can identify customer clusters $C_1, C_2, \dots, C_K$. For each service center s_j, the cluster C_j consists of the customers whose probability distributions $P_k(S_*)$ are located closest to the vertex π_j of the simplex Δ. We then proceed to the path optimization task, which we solve for each service center s_j with its customers from the set of locations C_j.

We also propose a method based on spectral properties of the graph. We apply multidimensional scaling (MDS) [14] to graph G to find the node, which has properties of the graph's center, and thus will minimize the average path to the rest of the nodes in the cluster C_j. After detecting such a node c, we check if any nodes adjacent to the center have a smaller average path. Recalculating the average path from neighboring nodes can be done with little effort. Instead of directly calculating all paths, we propose the following algorithm. Let v be one of the adjacent nodes connected to c by an edge e. Among the paths from the current center to all nodes, we can distinguish two types of ways: those that pass via the vertex v and those that do not. For paths that do have node v in their list of vertices, moving center from c to v reduces the path by the length (weight) of edge e, and for those that do not have v in their list, will increase paths length no more than by the length of e. Move the center to the vertex that has the minimum average path among the adjacent vertices, provided that it is less than the average path from the vertex c. We continue the iterative algorithm until we find a local minimum, when the average path in the neighboring vertices does not improve the situation.

The entire algorithm of network optimization can be described in the following pseudocode:

```
Data: Adjacency matrix of graph $G$,
      number of service centers $K$
Result: Location $S$ of the service network
Random initialization of $S \subset V$
while criterium of optimality is not achieved
    Construct the gravitational potential $\Psi_S$
    Construct the gradient flows (paths)
    Build the probability matrix $P(S)$
    Calculate the metric $J(S)$
    Find vertex $v_*$ with the highest potential value and
    change $S$ by moving one of the service centers to
    the vertex $v_*$ (the center that leads to the best value
    of the J-metric in the new configuration $S$)
end
Divide customers into clusters according to S
Perform path optimization in clusters
```

The algorithm performs $O(KN^3)$ operations at each while-loop iteration. Indeed, the most expensive is the search for eigenvalues and eigenvectors of the matrix L_S, which takes $O(N^3)$. The running times of the algorithms for constructing paths and the matrix P (S) are $O(N+M)$ and $O(N^2)$, respectively. Here M is the number of edges of G. The multiplier K appears before N^3 because the search for a service center to move to vertex v_* contains a hidden loop. When searching for an exact solution to the problem by the brute-force method, it is necessary to perform $\binom{N}{K}$ iterations.

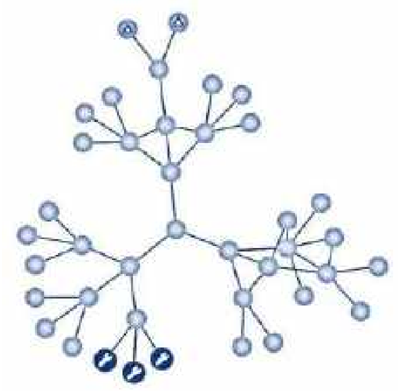


Fig. 4. Initial location of the service network.

Numerical experiments show that our greedy algorithm performs $O(K)$ iterations with high likelihood.

Assume that $K = \theta N$, where $\theta \in (0,1)$ is the share of locations where service centers need to be established. Then the total work of the brute-force method is

$$O\left(N^{3.5}\left(\frac{1}{\theta}\right)^{\theta N}\left(\frac{1}{1-\theta}\right)^{(1-\theta)N}\right).$$

Since both numbers θ^{-1} and $(1-\theta)^{-1}$ are greater than one, we are dealing with an NP-hard algorithm whose running time grows exponentially with the number of vertices in the graph. In contrast, the asymptotic running time of the greedy algorithm is only $cK^2N^3 = O(N^5)$.

V. Algorithm Performance

In conclusion, we will demonstrate the performance of the developed optimization algorithm on a simple graph. Consider the underlying graph formed from a balanced tree by adding or removing a few vertices and edges. We randomly selected weights from the range [0.5, 1] for the weighted graph. The graph still clearly shows three clusters, see Fig. 4. We initially located three service centers at the nodes marked by wrenches. This network configuration is far from optimal, especially since we have a bottleneck situation. The graph has 40 vertices. Thus, there are 9880 different service network locations that the brute-force method will analyze to find the exact solution.

The gravitational potential for the initial network configuration reaches its maximum value at the vertices marked with triangles. These vertices are located farthest from the set S. At the first stage of optimization, the influence of the weight matrix W is insignificant. The topology of the underlying graph influences the potential, and the "path length" at this stage can be interpreted as the number of edges that form the path. Let us move one of the service centers to a vertex with a triangle, see Fig. 5. In this configuration, it does not matter which of the centers is moved to which vertex.

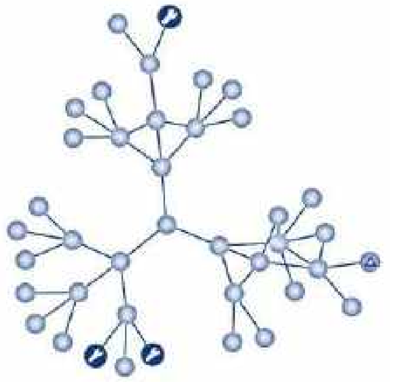


Fig. 5: Location of the service network after the first iteration.

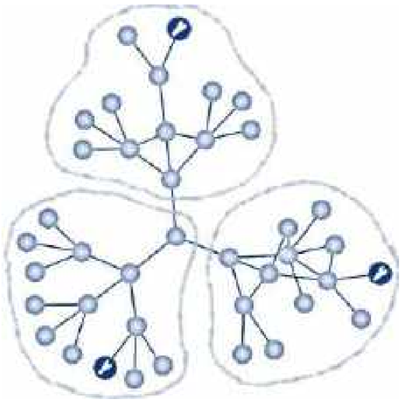


Fig. 6: The second iteration, where the global minimum of the optimization criterion is achieved.

Now the location farthest from the service network is in the cluster without a service center. The gravitational potential effectively finds places farther away from S, acquiring large values there. This feature of the potential is due to the variational properties of the eigenvalues and eigenvectors of the Laplace matrix. By moving one of the service centers to this vertex, we get the smallest possible value of the J-metric, which coincides with the value obtained by the brute-force method, see Fig. 6. After performing a few more control iterations and making sure that there is no improvement in the network configuration, we stop the computational process. Then each customer chooses the nearest service center, following the path of the maximum decrease in gravitational potential. Therefore, we get the clustering of the location graph.

In Fig. 7, we see the result of our algorithm after path optimization. In this example, our greedy algorithm achieved an accurate result when optimizing the J-metric. One reason is that the number of hidden clusters in the graph and service centers is the same, and all clusters are the same size. For example, the algorithm will work equally well for 6 or 9 centers. Our average path optimization algorithm also worked well here. The mean average precision error is only 3% compared to the result obtained by the brute-force method.

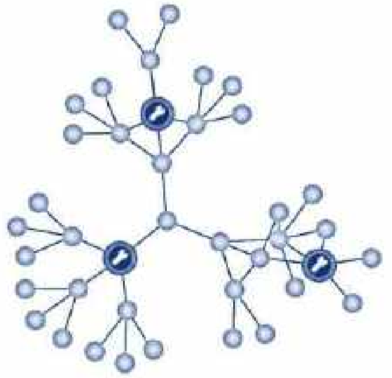


Fig. 7: Final configuration of the service network after optimizing average paths in clusters.

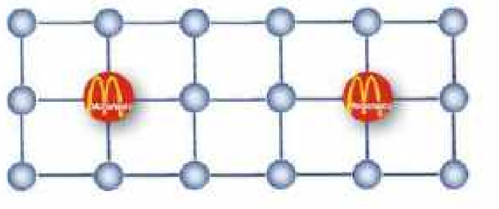


Fig. 8: The optimal service network in a grid graph.

Note that our visualization does not take into account the weights, i.e., the distance between nodes; this is the reason the location of the service center in the right cluster may not look optimal.

Experiments conducted on networks of various configurations and complexity showed high performance of the developed algorithm; they will be reported in detail in a future publication.

VI. Further Remarks

We note that our algorithm can be applied to the problem of the expansion of an existing service network $S_0 \subset V$. Suppose we must add new service centers $S = \{s_1, \dots, s_K\}$ to it so that the new network becomes as efficient as possible. Of course, $S \subset V \setminus S_0$. To solve the problem, we build a gravitational potential that vanishes at vertices from the set $S_0 \cup S$. However, we can move only service centers from S when optimizing the criteria.

Although our algorithm divides the graph into clusters, the problem of finding the optimal service network is not directly related to the graph clustering problem. The algorithm will efficiently place several restaurants in a city with perpendicular avenues and streets, even if the graph, which is now a square grid, has a homogeneous structure and does not contain any hidden clusters, see Fig. 8.

However, if there are such clusters in our graph, we believe that the method of gravitational potentials can be applied to the graph clustering problem as well. In Fig. 6, our algorithm found three clusters in the second iteration. Indeed, it is essential here that the number of service centers coincides with the number of clusters. Suppose graph G consists of m clusters with almost the same number of vertices. We experimentally found that the optimal network has the lowest J-metric values when K is a multiple of m. More precisely, if we consider the function $f(K) = J(S_*(K))$, where $S_*(K)$ is the optimal location of the network of size K, then f reaches local minima at points $K = km,\ k \in \mathbb{N}$. So, by applying our algorithm for different K, we can predict the number of clusters in G and find these clusters.

Conclusions

In this paper, we suggested an efficient algorithm of optimizing service centers location in a complex customer network that ensures balanced and predictable service load and proximity to the clients. The core of the method is the introduced concept of the gravitational potential that is based on the spectral properties of the graph Laplacian and allows one to find approximate shortest paths to closest servers. The proposed greedy algorithm then determines a predictable and balanced service area division among clients that optimizes nondeterministic customer preferences and the corresponding metrics. Finally, the iterative process starting from the MDS-inspired centroid optimizes the server location within each cluster. Numerical experiments (to be reported upon elsewhere) demonstrate that the developed pipeline is time efficient and returns nearly optimal solutions for all tested types of networks.

References

[1] S. Steinerberger, “A spectral approach to the shortest path problem,” Linear Algebra and its Applications, vol. 620, pp. 182–200, 2021.

[2] M. Newman, Networks. Oxford: Oxford University Press, 2018.

[3] E. Estrada, The structure of complex networks: theory and applications. Oxford: Oxford University Press, 2012.

[4] W. Kocay and D. L. Kreher, Graphs, algorithms, and optimization. CRC Press, 2016.

[5] F. Chung, Spectral graph theory. CBMS Regional Conference Series in Mathematics, vol. 92. Providence, Rhode Island: American Mathematical Soc., 1997.

[6] Handbook of cluster analysis, C. Hennig, M. Meila, F. Murtagh, R. Rocci, Eds. London: CRC press, 2015.

[7] M. C. Nascimento and A. C. De Carvalho, “Spectral methods for graph clustering – a survey,” European Journal of Operational Research, vol. 211(2), 221–231, June 2011.

[8] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, “A survey of kernel and spectral methods for clustering,” Pattern recognition, vol. 41(1), pp. 176–190, January 2008.

[9] M. Girvan and M. Newman, “Community structure in social and biological networks,” Proc. of the National Academy of Sciences, vol. 99(12), pp. 7821–7826, June 2002.

[10] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” Neural computation, vol. 15(6), pp. 1373–1396, June 2003.

[11] L. K. Saul, K. Q. Weinberger, F. Sha, J. Ham, and D. D. Lee, “Spectral methods for dimensionality reduction,” in Semi-supervised learning, O. Chapelle, B. Schölkopf, and A. Zien, Eds. Boston: MIT Press, 2006, pp. 292–308.

[12] D. F. Gleich and M. W. Mahoney, “Using local spectral methods to robustify graph-based learning algorithms,” Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 359–368, 2015.

[13] M. Zou, R. T. B. Ma, X. Wang and Y. Xu, “On Optimal Service Differentiation in Congested Network Markets,” in IEEE/ACM Transactions on Networking, vol. 26, no. 6, pp. 2693–2706, Dec. 2018.

[14] T. F. Cox and M. A. Cox, “Multidimensional scaling on a sphere,” Communications in Statistics – Theory and Methods, vol. 20(9), 2943–2953, 1991.

SIR Models on Complex Networks and Impact of Vaccination

Khrystyna Buhrii
Ivan Franko National University of Lviv
Lviv, Ukraine
khrystyna.buhrii@lnu.edu.ua

Yuriy Golovaty
Department of Mathematical Statistics and Differential Equations
Ivan Franko National University of Lviv
Lviv, Ukraine
yuriy.golovaty@lnu.edu.ua

***Abstract* — We study network models in epidemiology and different vaccination strategies to combat epidemics. The main goal is to compare the effectiveness of vaccination scenarios on complex networks with different topological structures: random, scale-free, and small-world networks. The results obtained during the study can be used to plan and evaluate disease control programs or to defend computer networks against a virus attack.**

***Keywords* — epidemiological model, SIR model, network model, random network, scale-free network, small world, vaccination, centrality measure.**

I. Introduction

Epidemiological models describe the spread of infectious diseases among human or animal populations. They allow us to track, predict, and inform epidemic response measures. Mathematical modeling of epidemics has a long history of research; see, for example, [1], [2] and references therein. Since the pioneering work of Kermack and McKendrick [3] in 1927, the so-called compartmental models have been extensively studied in the framework of the theory of dynamical systems. This approach involves dividing the population into different compartments based on disease status and monitoring the dynamics of each group. The natural history of illnesses reveals numerous epidemiological stages, such as susceptibility, resilience, incubation, illness, infectivity, etc. The most widely used SIR model categorizes individuals as Susceptible, Infectious, or Recovered. Various applications, not just epidemiology, caused considerable interest in these models. The compartmental models have been used to study sustainable agriculture, drug and alcohol use, the spread of violent ideologies on the internet, and criminal activity [4].

The emergence of the internet and computer viruses sparked a new wave of research in the field of epidemiological models. This occurred in the late 20th century when the spread of computer viruses began to pose a serious threat to computer system security. The application of epidemiological models to computer networks has required careful adaptation because the spread of computer viruses is not entirely analogous to the spread of biological diseases. What happens to a computer virus in a network can only be adequately described by considering the network topology and various network characteristics. That is why a new type of epidemiological model, called network models, has emerged. These models analyze the structure of computer networks and the patterns of communication between nodes. By studying the network topology and data flow between systems, these models can identify critical nodes or vulnerable areas where viruses can spread rapidly [5]. Network models can also be used to evaluate the effectiveness of network segmentation, firewalls, and other security measures and even to detect the sources of computer viruses in networks [6]. These models effectively describe various processes for which interconnections are essential. In economics, these include job referrals in labor markets, patterns of international trade, the diffusion of technology, and contagion in financial markets [7].

With the Covid-19 pandemic sweeping the globe, the network models developed for computer networks have come back in epidemiology. Only they allow us to study the effect of vaccination and develop an optimal vaccination strategy, especially in conditions of limited vaccine supplies. In compartmental models, we assume everyone has an equal chance of getting sick. However, this assumption does not reflect that people interact in much smaller groups. Network models consider real-world interaction patterns, provide valuable information about the spread of infectious diseases, and effectively suggest ways to combat them [5] - [8].

We investigate the relationship between the topological characteristics of complex networks where the virus spreads and the effectiveness of different vaccination scenarios. Real-world networks share a few common properties, such as the small-world and scale-free phenomena. In addition, they are sparse, as the fraction of links is relatively small compared to the links in the complete graph. We build probabilistic models of epidemic spread and localization through vaccination for the cases when one of these properties is crucial in a network. Vaccination scenarios depend on different centrality metrics for graphs. We performed statistical modeling to compare the vaccination efficiency for different scenarios and network structures. Vaccines have saved more lives than any other medical breakthrough in history. Therefore, identifying the most vulnerable individuals and their vaccination plays an essential role in fighting epidemics, which is why it is an essential topic for the studies.

II. SIR Models on Complex Networks

Let us consider a population of individuals in contact with each other. At some point, a random individual contracts an infectious disease that spreads through the population. After some time, the share of patients will reach a detectable level, which we call the threshold. Then the infection begins to threaten the population, and measures, such as vaccination or quarantine, must be taken to stop the epidemic.

According to the ideology of compartmental models, we divide the population into three non-intersecting groups: S (susceptible), I (infectious), and R (recovered), see Fig.1. Let us assume that $S(t), I(t)$, and $R(t)$ are functions of time that describe the size of the groups. Individuals can move between

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

Susceptible
Infectious
Recovered

Fig. 1: States in the SIR epidemic model and transitions between them.

these compartments according to the scheme shown in Fig. 1. In the classic SIR model, the dynamic system

$$\frac{dS}{dt} = -\frac{\beta}{N}SI, \qquad \frac{dI}{dt} = \frac{\beta}{N}SI - \gamma I, \qquad \frac{dR}{dt} = \gamma I$$

describes the evolution of the groups. Here N is the population size, β is the average number of contacts per person per time and γ is the average number of recovered infectious individuals. We assume that $S(0) = N, I(0) = 0, R(0) = 0$. The typical behavior of solutions of the dynamic system is shown in Fig. 2. In this model, $S(t) + I(t) + R(t) = N$ for all t. The epidemic lasts as long as the share of infected people exceeds the threshold θ, i.e., $I(t) > \theta N$.

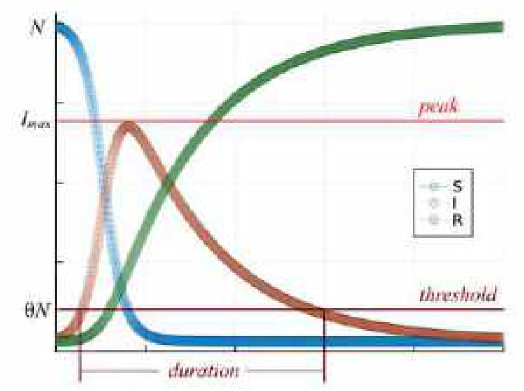


Fig. 2: Dynamics of compartments in the SIR model
© computationalthinking.mit.edu

The critical indicators of an epidemic are its duration and the peak number of infected people. Long-lasting epidemics and pandemics have many negative economic and social consequences, and the number of deaths is proportional to the peak value I_{max}. Both indicators can be reduced by introducing quarantine, which lowers the intensity of contact among the population, and vaccination, which significantly decreases the chances of getting sick. The classical SIR model can respond to such measures only after the fact, reducing the parameter β. A strict limitation of the model is the assumption that all groups are uniformly mixed in the population and that everyone has the same probability of infection. It does not provide mechanisms for analyzing any external influence on the epidemic. However, it is well known that "social interactions are not organized in this stylized way. Instead, individuals interact mostly within much narrower groups, shaped, for example, by family ties, work and social environments, and geography. Network models provide a route into analyzing epidemics in a way that takes these patterns of interaction into account" [7].

Let us denote by $G = (V, E)$ the connected undirected graph, where V is the set of vertices, and E is the set of edges. This graph models a population with all its members and the connections between them. A vertex is an individual, and an edge indicates that two individuals are in contact with each other. Let us divide V into three disjoint subsets S, I, and R, assigning each vertex one of the attributes – susceptible, infectious, or recovered. The epidemic dynamics consists of the change of these subsets in time, considered discrete in the network model. Graph G is the stage where we will examine several cases of epidemic development and its control. The cases will differ in topological characteristics of G and vaccination strategies, which are also related to specific graph metrics.

Assume $I(t)$ is not empty, and describe what happens to the population after the passage

$$[S(t), I(t), R(t)] \rightarrow [S(t+1), I(t+1), R(t+1)]. \quad \text{(T)}$$

Only vertices from $I(t)$ can cause changes in the groups. Each vertex $u \in I(t)$ can move to $R(t+1)$ with probability γ, i.e., change the label from *'susceptible'* to *'recovered'*. Alternatively, u can infect one of its neighbors. A neighbor is any vertex that is connected to u by an edge. We randomly chose a neighbor v from $S(t)$, and then we will move this vertex from $S(t)$ to $I(t+1)$ with probability β, replacing the label *'susceptible'* with *'infectious'*. We denote the transformation (T) of the labeled graph G by Spread(G, β, γ). The dynamics of the compartments on G can be obtained by iterating of Spread.

Let us assume that the population reached the epidemiological threshold at moment t_*. In our model, the function Vaccination(C) works as follows. We have selected a set of nodes $H \subset V$ according to a specific scenario C developed in advance. If $v \in H \cap S(t_*)$, then we assign the label *'recovered'* to this vertex. Other nodes from H remain in their groups. The infected nodes $v \in H \cap I(t_*)$ remain infected because vaccination is not a cure. Obviously, the recovered nodes $v \in H \cap R(t_*)$ remain recovered.

We simulate the spread of an epidemic in a complex network to calculate the following parameters:

- *Epidemic duration:* the number of iterations of Spread during the epidemic lifecycle, i.e., from the beginning to the recovery of the last patient.
- *Epidemic peak:* the largest number of infected individuals simultaneously, i.e., $I_{max} = \max_t |I(t)|$.
- *Epidemic coverage:* the total number of individuals infected during the epidemic lifecycle.

The spread of the epidemic in our population occurs according to the following algorithm.

```
Input Data: network $G$, probability $\beta$ of infection, probability
$\gamma$ for the patient to recover, vaccination scenario $C$

Initialize all vertices as susceptible
Initialize a random vertex as infectious

# Spread of infection before the threshold is reached
while $|I(t)| < \theta|V|$
    Spread$(G, \beta, \gamma)$
end

# Vaccination and spread of the epidemic until it is complete
Vaccination$(C)$
while $I(t)$ is not empty
    Spread$(G, \beta, \gamma)$
end

Output Data: duration, peak, coverage
```

III. Sample Networks

In this section, we describe the network structure of our populations that will have to survive an epidemic. The sample graph should resemble real-world networks or at least have their main features. In Fig. 3, we visually classify networks in the space of variables *randomness-heterogeneity-modularity*. The randomness dimension represents the number of randomnesses involved in the network construction process, the heterogeneity measures how diverse the distribution of connections is, and the modularity measures how modular the architecture is [9]. We can see that most real-world networks are in the domain of highly heterogeneous, random hierarchical networks.

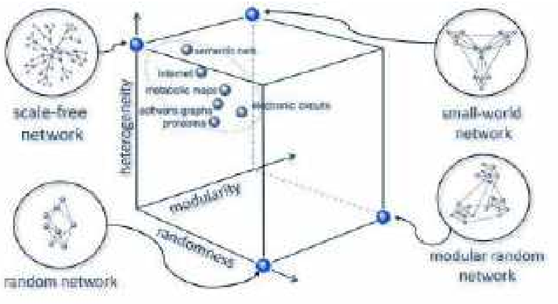


Fig. 3: A zoo of complex networks. © Solé and Valverde [9]

We used four types of graph structures: random networks, modular random networks, scale-free networks, and small-world networks. Each of the structures emphasizes particular features of real-world networks, see Fig. 3.

A. Random networks

Let us assume that the target population can be represented as a random network since contacts and disease transmission between people are usually random. The random networks were first introduced by Erdős and Rényi [10]. Due to the outstanding contributions of these mathematicians, random networks are commonly referred to as Erdős-Rényi graphs. We applied the following algorithm to create a random graph [11]. Among N isolated vertices, we select a pair of distinct vertices and connect them with probability p or skip them with probability $1 - p$. This process is repeated for each possible pair of vertices in the graph. The resulting distribution of vertex degrees follows the binomial distribution. Therefore, the degrees lie in a close neighborhood of some mean degree, and there are usually no outliers, i.e., vertices with very high or low degrees. We generated a giant connected component of a random graph with parameters $N = 300$ and $p = 0.1$ using *NetworkX,* the Python package for creating and manipulating graphs. From now on, this network will be regarded as a *random network* instead of a *modular random network*, which we will describe below.

The population can often be divided into relatively isolated groups or clusters. For example, when we consider people from different cities, it is reasonable to assume that most connections are within their city, with a limited number of people communicating between the cities. A graph representing this population should have a more clustered structure. We create a modular random network as follows. Firstly, we generate three random graphs of size $N = 100$, then select a few nodes from each cluster and connect them to some nodes from two other ones. As a result, we will receive a random network of size N = 300 consisting of three clusters.

B. Scale-free networks

As we mentioned above, many real-world networks are scale-free. A scale-free network is a graph whose degree distribution follows a power law, meaning there are large hubs – nodes with much higher degrees than most other nodes in the network. The Barabási-Albert model is the best-known model for creating a scale-free network that connects new nodes to an existing graph by combining two concepts: growth and preferential attachment [11].

We can build a scale-free graph using the following algorithm. We start with m_0 vertices, the connections between which are chosen randomly so that each vertex has at least one connection. The network develops in two steps.

- *Growth.* At each iteration we add a new node with $m \leq m_0$ links that connect the new node to m nodes that are already in the network.
- *Preferential attachment.* The probability that a link of the new node connects to a node v depends on the degree of v. Namely, nodes with a higher degree have a stronger ability to grab links added to the network.

As for a modular random network, we first generate three Barabási-Albert clusters with $N = 100$ and $m = 3$, and then combine them into a single graph of size 300, see Fig. 4.

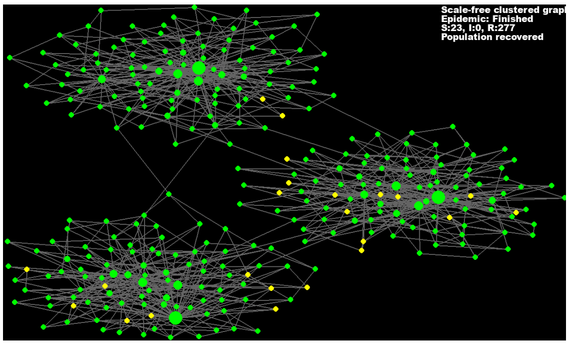


Fig. 4: A sample of a scale-free network

C. Small-world networks

In a small-world network, the average length of the shortest path connecting two nodes grows very slowly as the network size increases. Watts and Strogatz proposed a model that generates graphs with small-world properties [12], [13]. We start with a lattice (e.g., a ring) of N nodes connected to $2m$ nearest neighbors. We iterate over the nodes, and at each step, with probability p, one link connecting the selected node to one of its m nearest neighbors is reconnected to another randomly selected node, and with probability $1 - p$, all links remain in place. Self-connections and duplicate edges are excluded.

Next, by modeling three relatively isolated communities, we generate three small worlds with parameters $N = 100$ and $p = 0.5$. Then we connect them with a small number of random edges into a single network.

All generated networks are the same size to compare an epidemic's spread statistics. Note that the classical SIR model, in which all three groups are uniformly mixed, and each member of the population can get sick with equal probability, has an analog in network models. To do this, we need to consider the epidemic on the complete graph.

IV. Vaccination Scenarios and Results of Simulations

Our population $G = (V, E)$ consists of a finite number $|V|$ of individuals. Since the probability γ is positive, all members from compartment I will move to compartment R in a finite time, i.e., all will recover. Therefore, by repeatedly applying the transformation Spread, we will come to a labeled graph with $|I(T)| = 0$ for a finite number of iterations T. The time T is the duration of the epidemic.

The approach in which we do not interfere with the spread of the epidemic and do not take any measures to stop it is unacceptable for many infectious diseases. Such inaction can have many negative economic and social consequences and lead to high mortality among the population. The epidemic can be influenced by quarantine and vaccination of the population. These measures can shorten the duration of the disease and reduce the peak incidence.

Below we will describe several vaccination methods to compare their effectiveness. The vaccination scenarios, except for random vaccination, are based on selecting important nodes in the network. Centrality metrics on graphs will determine this importance [14], [15].

We modeled the spread of the epidemic with parameters $\beta = 0.7, \gamma = 0.1$, and $\theta = 0.05$ for a population of size $N = 300$. In each case, except for the "No Vaccination" scenario, we used 15 vaccine doses, meaning that only 5% of the population was vaccinated. We simulated the epidemic in four different networks with six different vaccination scenarios. We ran the simulation 1000 times for each case to account for random effects and then calculated the average values of the epidemic's duration, peak, and coverage.

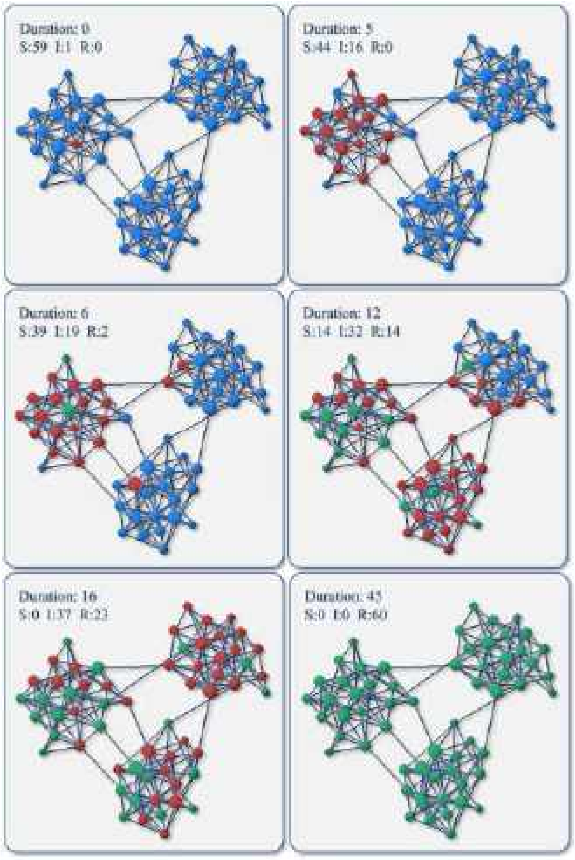


Fig. 5: Frames of the epidemic visualization on a modular random network of size N=60.

The results of our simulations are tables with statistical indicators and visualization of the actual dynamics of the epidemic in all types of networks and for all vaccination scenarios. The reader can watch these videos at linktr.ee/IEEE.ELIT.2023 [16].

In Figure 5, we visualize only the main stages of an epidemic in a small network. Initially, the disease spread to people in one community, but on the sixth day, the epidemic spread to other communities. On the 12th day, two communities were already affected, but the situation was manageable in the third. However, everyone later contracted the disease, and the epidemic ended only on the 45th day.

Let us take a closer look at all the methods of vaccination and present the statistics from our simulations.

TABLE I. No vaccination scenario

Network	*Duration*	*Peak %*	*Coverage %*
Random graph	71	81	100
Modular random graph	76	54	100
Scale-free graph	78	46	97
Small-world graph	79	47	99

No Vaccination. In this scenario, we do not use vaccines to track the dynamics of the disease without external intervention and compare it with the effect of vaccination.

In the tables, we provide the duration of the epidemic in terms of the number of iterations and the peak and coverage of the disease as a percentage of the population.

TABLE II. Random vaccination scenario

Network	*Duration*	*Peak %*	*Coverage %*
Random graph	70	77	95
Modular random graph	76	51	94
Scale-free graph	78	42	92
Small-world graph	78	43	93

o *Random Vaccination.* Let us randomly select 5% of people and vaccinate them. This scenario is the simplest method of vaccination. However, even it has the effect of slightly reducing the peak of the disease.

TABLE III. Degree vaccination scenario

Network	*Duration*	*Peak %*	*Coverage %*
Random graph	70	77	95
Modular random graph	76	50	94
Scale-free graph	78	39	89
Small-world graph	80	36	88

o *Degree Vaccination.* In this scenario, the degrees of the network nodes are essential. Nodes with the highest degrees are considered central to the network. This metric measures the intensity of direct contact between an individual and his or her neighbors. Individuals in contact with the largest number of others have the highest centrality score. After sorting the list of vertices in descending order for degree, we select 15 individuals from the top of the list and vaccinate them.

○ *Closeness Vaccination.* Unlike the previous method, which relied on the social importance of people and the intensity of their contacts, this vaccination scenario is based on their "geographic" location. The closeness measure for a node depends on the average distance from the node to all other nodes in the network. The node is essential if we can get from it to any other node in smaller steps. Let d_{uv} denote the distance, in the network G, between vertices u and v measured as the minimum numbers of hops needed to move from u to v. The mean distance from u to any other node is given by

TABLE IV. CLOSENESS VACCINATION SCENARIO

Network	*Duration*	*Peak %*	*Coverage %*
Random graph	71	77	95
Modular random graph	64	29	47
Scale-free graph	78	33	81
Small world	78	29	71

$$d_u = \frac{1}{|V|-1}\sum_{v\in V} d_{uv}.$$

If a vertex u has small d_u, then it is close to many nodes of the network. We call $C_u = {d_u}^{-1}$ the closeness centrality of u. In this vaccination method, we sort the array $(C_u)_{u\in V}$, select 15 individuals with the highest closeness centrality, and vaccinate them.

○ *Betweenness Vaccination.* The measure of betweenness calculates how often the node is found on the shortest path between two random nodes of the network. This metric states that the node is important if it is a kind of a gateway for the network. Such nodes are often located in network bottlenecks. In Figures 4 and 5, they correspond to those members of the population who have contacts with representatives of other communities, i.e., they are the ends of edges connecting different clusters.

TABLE V. BETWEENNESS VACCINATION SCENARIO

Network	*Duration*	*Peak %*	*Coverage %*
Random graph	71	77	95
Modular random graph	58	25	33
Scale-free graph	79	34	81
Small world	66	24	48

The betweenness centrality B_u for a node u is defined as follows. Let s_{vw} be the number of shortest paths between vertices v and w in the network G, and m^u_{vw} be the number of shortest paths between vertices v and w that pass through node u. The betweenness centrality B_u of u is the value

$$B_u = \sum_{v,w\in V}\frac{m^u_{vw}}{s_{vw}}.$$

We select the 5% of individuals with the highest values of the betweenness centrality and vaccinate them.

○ *Eigenvector Vaccination.* Such measures as the degree, closeness, and betweenness centralities can identify people in the population who may be active spreaders of infection. While the eigenvector centrality identifies the most at-risk group in the population, the individuals most likely to become infected. The eigenvector centrality gives each node a score proportional to the sum of the scores of its neighbors. This principle is used in the PageRank algorithm to rank web pages. In the context of our model, we can reformulate it as follows. An individual's infection probability at the beginning of the epidemic is proportional to the sum of the probabilities of infection of all those with whom it comes into contact.

TABLE VI. EIGENVECTOR VACCINATION SCENARIO

Network	*Duration*	*Peak %*	*Coverage %*
Random graph	71	77	95
Modular random graph	77	48	93
Scale-free graph	79	39	89
Small-world graph	79	41	92

Such a recursive definition of this measure leads to the spectral problem $Ax = \lambda_{max}x$, where A is the adjacency matrix of G, and λ_{max} is the leading eigenvalue of A. The normalized eigenvector x can be chosen to be positive. The vector gives us the eigenvector centrality score, and its coordinates are the probabilities of infection for the corresponding vertices, i.e., if the jth row of A corresponds to vertex u, then x_j is the probability of getting sick for this vertex.

After calculating the eigenvector x, we vaccinate 15 individuals, corresponding to the nodes with the highest score.

V. SOME REMARKS ON QUARANTINE

How can network models consider the impact of quarantine on the epidemic's dynamics? Quarantine only affects the intensity of contact among the population. Let p be the probability of a healthy person becoming infected with the virus when meeting an infected person. If the population members have the same immunity, this probability depends only on the biological properties of the virus, such as the rate of reproduction or mode of transmission. Next, if two individuals connect, assuming they are in constant contact is unnecessary. Suppose an edge with a weight w_{uv} connects vertices u and v. We can interpret w_{uv} as the probability of contact. The person u can transmit the virus to person v with probability p if they meet, which will happen with probability w_{uv}. Hence the transmission probability is $\beta_{uv} = pw_{uv}$.

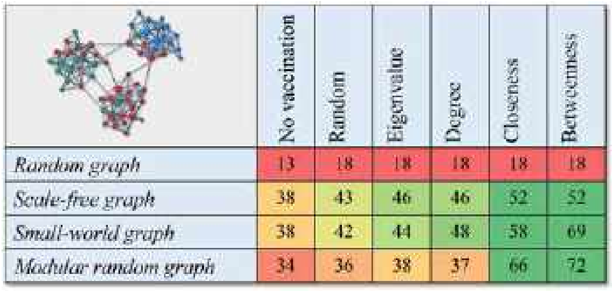
Let $G_t = (V, W_t)$ be a sequence of weighted graphs with matrix $W_t = (w_{uv}(t))_{u,v\in V}$. Then the infection spread rate $\beta_{uv}(t) = pw_{uv}(t)$ depends on time and a pair of individuals. Suppose we have reduced the entries of W_t by a percentage during a certain period $[t_{1,}t_2]$. In that case, this means a decrease in the intensity of contacts in the population, i.e., the introduction of quarantine in the entire population. Reducing only a part of the probabilities $w_{uv}(t)$ for a certain time is also possible, which would mean local quarantine, for example, in only one community. The impact of different quarantine strategies in epidemiologic models still needs to be studied.

CONCLUSIONS

Our modeling aims to answer the question of who should be vaccinated first at the beginning of the epidemic, thereby reducing its negative consequences. It is impossible to vaccinate large groups of people quickly, so vaccinating only 5% of the population in our model is not only due to a limited number of vaccines. Statistical tests show that any vaccination scenario affects the spread of the epidemic by reducing the peak incidence. In addition, vaccination strategies, based on network structure analysis are more effective than random vaccination; see Tables I-VI. However, the level of such effectiveness varies. Table VII shows an integral indicator of vaccination efficiency in different networks. It takes into account both the peak and the coverage of the epidemic.

First, the epidemic simulation on a giant connected component of a random graph indicates that vaccination is ineffiient, regardless of the strategy. The network must be highly modular to achieve a visible result by vaccinating a small percentage of the population.

TABLE VII. VACCINATION EFFECTIVENESS ACCORDING TO THE INTEGRAL SCORE $\varepsilon = 100 - (0.7 \cdot \text{PEAK} + 0.3 \cdot \text{COVERAGE})$



	No vaccination	Random	Eigenvalue	Degree	Closeness	Betweenness
Random graph	13	18	18	18	18	18
Scale-free graph	38	43	46	46	52	52
Small-world graph	38	42	44	48	58	69
Modular random graph	34	36	38	37	66	72

The betweenness and closeness methods show the best results in clustered networks; see Tables V-VII. These metrics select people who communicate the most with members of other communities. In other words, nodes with high betweenness or closeness scores can usually be found on bridges between clusters. Vaccinating such individuals prevents the disease's spread by blocking the infection pathways between relatively isolated groups. In one of the videos available to the reader on the website [16], vaccination with the betweenness method led to the localization of the epidemic within only one cluster.

The spectral method, which is based on the eigenvector centrality, is less effective in the case of fast vaccination at the beginning of the epidemic. It selects a group with a high disease risk, usually comprising a significant portion. Vaccinating people in this group does not affect the global spread of the disease. This method is better suited for systematic vaccination in a prolonged pandemic. Also, all the vaccination scenarios based on the centrality metrics gave worse results on the scale-free network compared to the small-world and modular random networks. The structure of the scale-free graph can explain this. In Figure 4, the size of the nodes is shown according to their degree so that the size is larger for nodes with higher degrees. It is easy to see that, in the case of the scale-free network, the nodes with high centrality scores are the nodes with high degrees located in the centers of clusters. "In the centers" means far from the gateways or bridges that connect these clusters. There is no doubt that vaccinating such individuals with many contacts is useful. However, it is less effective than vaccinating people with inter-cluster connections.

The duration of a virtual epidemic is almost independent of network topology and vaccination strategies. For a network of size $N = 300$, this parameter generally ranges from 70 to 80. In addition, the standard deviation is also almost the same and ranges from 11 to 13. The fact is that the duration of an epidemic depends mainly on only three parameters: population size N, contact rate β and recovery rate γ. When β and γ are fixed, the duration can be expressed by the formula $D = a + b \ln N$, where the coefficients can be found using simple linear regression. The dependence between duration and the other parameters is more complex and advanced methods such as neural networks should be used to properly analyze them.

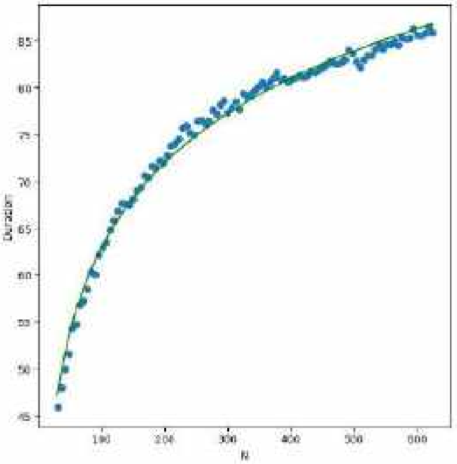


Fig. 6: Duration based on network size with other parameters fixed: $\beta = 0.7$ and $\gamma = 0.1$. $D = 3.02 + 13\, ln\, N$.

REFERENCES

[1] Brauer, F., van den Driessche, P., and J. Wu. Mathematical Epidemiology. Springer Science and Business Media, 2008.

[2] Lypez-Flores, M. M., Marchesin, D., Matos, V., & Schecter, S. Differential Equation Models in Epidemiology, 2021.

[3] Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London. Series A,, 115(772), 700-721.

[4] Koss, L. (2019). SIR models: differential equations that support the common good. CODEE Journal, 12(1), 6.

[5] Ganesh, A., Massoulié, L., & Towsley, D. (2005, March). The effect of network topology on the spread of epidemics. In Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies. (Vol. 2, pp. 1455-1466). IEEE.

[6] Shah, D., & Zaman, T. (2010). Detecting sources of computer viruses in networks: theory and experiment. In Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems (pp. 203-214).

[7] Craig, B. R., Phelan, T., Siedlarek, J. P., & Steinberg, J. (2020). Improving epidemic modeling with networks. Economic Commentary, (2020-23).

[8] Witten, G., & Poulter, G. (2007). Simulations of infectious diseases on networks. Computers in Biology and Medicine, 37(2), 195-205.

[9] Solé, R. V., & Valverde, S. (2004). Information theory of complex networks: on evolution and architectural constraints. In Complex networks (pp. 189-207). Berlin, Heidelberg: Springer Berlin Heidelberg.

[10] Erdos, P., & Renyi, A. (1960). On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci, 5(1), 17-60.

[11] Posfai, M., & Barabasi, A. L. (2016). Random Networks. Network Science. Cambridge, UK::Cambridge University Press.

[12] Barrat, A., & Weigt, M. (2000). On the properties of small-world network models. The European Physical Journal B-Condensed Matter and Complex Systems, 13, 547-560.

[13] Albert, R., & Barabasi, A. L. (2002). Statistical mechanics of complex networks. Reviews of modern physics, 74(1), 47.

[14] Njotto, L. L. (2018). Centrality Measures Based on Matrix Functions. Open Journal of Discrete Mathematics, 8(04), 79.

[15] Latora, V., Nicosia, V., & Russo, G. (2017). Centrality measures. Complex Networks, 31-68.

[16] Buhrii, K. Simulation of epidemic spread. linktr.ee/IEEE.ELIT.2023

Optimizing Neural Network Wavefunctions Using Variational Monte Carlo with Evolution Strategies

Mykhailo Moroz
Department of Solid State Physics
Ivan Franko National University of Lviv
Lviv, Ukraine
michael08840884@gmail.com
mykhailo.moroz@lnu.edu.ua

Oleg Bovgyra
Department of Solid State Physics
Ivan Franko National University of Lviv
Lviv, Ukraine
bovgyra@gmail.com
oleh.bovhyra@lnu.edu.ua

Abstract — **Obtaining accurate solutions to many-electron quantum systems is of critical importance, in principle, such solutions allow computing most of possible chemical and physical properties from first principles. Exact solutions require utilizing exponentially scaling algorithms, which are not practical for large systems. Neural networks provide a new approach to approximating the solution while keeping a high level of accuracy. In this work, we explore a novel method of optimizing neural network wavefunctions by employing Evolution Strategies (ES) within a Variational Monte Carlo (VMC) framework. This allows us to more robustly find solutions while keeping a small number of parameters, making it an efficient alternative to traditional methods.**

Keywords — Neural Network Wavefunction, Variational Monte Carlo, Evolution Strategies, Machine Learning, Ab Initio solutions

I. Introduction

The explosion of research in Machine Learning has led to interesting new ways to utilize neural networks. Their ability as general function approximators has firmly placed them as an invaluable tool when it is difficult to directly describe a functional relationship. One of such promising new directions has emerged in the field of Numerical Quantum Chemistry, neural networks have shown themselves to be a great way to encode the properties of the system. The approaches can be broadly categorized into Supervised and Self-generative, in the former the neural network is fitted to replicate properties given from a database of precomputed chemical systems or crystals, such as potential-energy surfaces given the nuclei coordinates. The first of such neural networks was proposed by Behler et al. [1]. While it was significantly faster than Density Functional Theory (DFT) calculations, this had the natural drawback of the need to utilize other ab initio methods to generate the fitting data, and the generalizability as well as accuracy will in the end be limited by the size of the set of computed systems and the accuracy of the chosen base method. Our previous work focused on predicting such potential-energy surfaces for ZnO systems, the generation of the dataset for our model took days and any mistake in the generation increased the amount of time it takes to get a trained model [2]. In the case a mistake in the dataset was missed the final trained model might exhibit non-physical behaviour thus rendering the computed results not useful. On the other hand, self-generative methods try to use neural networks as the backbone of ab initio calculations of system properties, and as such we can take advantage of the flexibility of their representations.

Classically, to describe systems which can represent electron correlations to a high degree of accuracy, we can no longer rely on methods like Hartree-Fock or DFT, so methods like Coupled Cluster Single-Double (CCSD) with single or triple excitations, or full configuration interaction (FCI) can be used [3]. The issue in using such methods is their scaling with the system size, CCSD already requires $O(N^6)$ and FCI scales exponentially $O(N!)$ which limits their use to only very small quantum systems. It has been shown by Hermann et al., that neural networks can be used to represent the full wavefunction of the system instead, their flexibility to represent nonlinear features makes allows us to accurately describe the wavefunction without the need for such a large number of parameters [4]. As such neural networks promise to provide solutions at the cost of methods like HF, while keeping the accuracy of the more advanced methods.

II. The wavefunction neural network

We are mainly interested in describing wavefunctions representing many-body electron systems representing chemical bonds, this puts some constraints on the functional form of the wavefunction. To achieve as high accuracy as possible our architecture must conform to these requirements. To satisfy the Fermi-Dirac statistics we need the wavefunction ansatz to be antisymmetric under the exchange of two electron coordinates i.e.,

$$\psi_\theta(\ldots, \boldsymbol{x}, \ldots, \boldsymbol{y}, \ldots) = - \psi_\theta(\ldots, \boldsymbol{y}, \ldots, \boldsymbol{x}, \ldots) \qquad (1)$$

Pfau et al. [4] have shown that one can use a Slater-Jastrow like wavefunction ansatz where the one-electron wavefunctions are functions of the electron coordinates and are symmetric under the permutation of all electrons except the given one which are represented using feed-forward neural networks, with the symmetric part being a sum over contributions of different electrons [5]. They have named this architecture FermiNet. Allowing the one-electron wavefunctions to have this additional degree of freedom keeps the anti-symmetry of the Slater determinant while making the approximation capable of much more flexible wavefunction representations.

$$\psi_\theta(\boldsymbol{x}) = \exp(J_\theta(\boldsymbol{x}))\boldsymbol{\Sigma}^{\mathbf{N}}_{\mathbf{k=1}}\det(\phi_{\theta,k}{}^{\mathrm{down}}(\boldsymbol{x}))\det(\phi_{\theta,k}{}^{\mathrm{up}}(\boldsymbol{x})) \qquad (2)$$

Where $\exp(J_\theta(\mathbf{x}))$ is the Jastrow factor, θ is the set of all neural network parameters, $\phi_{\theta,k}{}^{\mathrm{down}}$ is the set of spin up electron orbitals and $\phi_{\theta,k}{}^{\mathrm{up}}$ is the set of spin up electron orbitals, both of which are described by the a neural network which is symmetric under exchange of other electron coordinates. The core part of the neural network is a feed-forward part consisting of n layers [6]. Such a network can be written down like this:

$$\mathrm{NN}(\mathbf{x}) = \sigma(\mathbf{B}_n + \mathbf{W}_n\sigma(\ldots\ \sigma(\mathbf{B}_1 + \mathbf{W}_1\boldsymbol{x})\ \ldots)) \qquad (3)$$

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

Where $\boldsymbol{x}$ is the vector of our input features, $\mathbf{B}_n$ and $\mathbf{W}_n$ are the bias vector and weight matrix respectively, and σ is the activation function, in our case the sigmoid function. To properly describe the wavefunction cusp, so that our ansatz satisfies the Kato cusp condition we need to add the distance to the nuclei as an input feature [7]. We shall use a concatenated vector of electron positions and distances relative to each nucleus.

$$\boldsymbol{x}_i = \{\boldsymbol{r}_i - \mathbf{R}_j, |\boldsymbol{r}_i - \mathbf{R}_j|\} \tag{4}$$

To build a function that is symmetric under electron exchange we can do a sum over all single-electron feature vectors and add contribution of the i-th electron separately.

$$\phi_{\theta,i}(\boldsymbol{x}) = NN_1(NN_2(\boldsymbol{x}_i) + NN_3(\boldsymbol{\Sigma}^N_{\mathbf{k=1}}NN_4(\boldsymbol{x}_k)) \tag{5}$$

The sum over all k-th electrons makes this symmetric under their exchange. NN_i are all separate neural networks with different specifiable input and output vector dimensions, but the input dimensions of the second and fourth neural networks are $4N_{nuclei}$ while the output dimension of the first neural network is equal to the number of electrons times the number of determinants $N_{det}n$, since we need n orbitals per electron per determinant. The output dimension of the second and third networks must coincide.

III. Finding the ground state of a system

To find the ground state wavefunction we need to find the lowest energy eigenvalue and eigenfunction of the Schrodinger equation, which can be written like this for the many-body electron case around a set of static nuclei under the Born-Oppenheimer approximation in Hartree units:

$$\widehat{H}\psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = E\psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \tag{6}$$

$$\widehat{H} = \sum_i^n \nabla_i^2 + \sum_{i>j}^n \frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} - \sum_{i,I}^{n,N} \frac{Z_I}{|\boldsymbol{r}_i - \boldsymbol{R}_I|} + \sum_{I>J}^N \frac{Z_I Z_J}{|\boldsymbol{R}_I - \boldsymbol{R}_J|} \tag{7}$$

Where $\psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ is the electron wavefunction for n electrons, $\boldsymbol{x}_i = \{\boldsymbol{r}_i, \sigma_i\}$, where $\boldsymbol{r}_i$ is the electron position and σ_i is the electron spin. ∇^2_i is the Laplacian with respect to $\boldsymbol{r}_i$, N is the number of nuclei, also $\boldsymbol{R}_I$ and Z_I are the nuclei positions and charges.

To find the set lowest energy solution of this equation we need to optimize the neural network parameters with the energy as the loss function. Since our wavefunction ansatz is not normalized to compute the energy we can use the Rayleigh quotient. Optimizing in such a way can be done by Variational Monte Carlo [8].

$$\mathcal{L}_\theta = \frac{\langle \psi_\theta H \psi_\theta \rangle}{\langle \psi_\theta \psi_\theta \rangle} \tag{8}$$

Because the wavefunction we are dealing with is high-dimensional, with 3N dimensions in total, the only practical approach of evaluation is by utilizing Monte Carlo methods. Even so naive sampling of the wavefunction would have been exceedingly ineffective, so we need to sample the integrals in regions with higher electron probabilities. To do so we can utilize the Metropolis-Hastings algorithm [9]. This algorithm is a variant of a Markov Chain Monte Carlo algorithm, where the chain of samples is correlated with each other. To avoid the correlation to have an impact on the integral estimate we must keep the sample chains well-mixed, such that each Markov Chain step has a sufficiently high chance of accepting this sample.

Rewriting the energy equation into integral form can substitute the equation for importance sampling proportionally to the electron probability $\psi^2(\boldsymbol{x})$ we get a division by it both in the numerator and denominator. As such we can remove the probability integral all-together and only keep $\psi(\boldsymbol{x})H\psi(\mathbf{x})/\psi^2(\boldsymbol{x})$

$$\mathcal{L}_\theta = \frac{\int d\boldsymbol{x}\psi_\theta(\boldsymbol{x})\widehat{H}\psi_\theta(\boldsymbol{x})}{\int d\boldsymbol{x}\psi_\theta(\boldsymbol{x})\psi_\theta(\boldsymbol{x})} = \boldsymbol{E}_{\boldsymbol{x}\sim\psi_\theta^2}\left(\psi_\theta^{-1}(\boldsymbol{x})\widehat{H}\psi_\theta(\boldsymbol{x})\right) = \boldsymbol{E}_{\boldsymbol{x}\sim\psi_\theta^2}\left(E_L(\boldsymbol{x})\right) \tag{9}$$

where the function $E_L(\mathbf{x})$ is known as local energy and is equal to $\psi(\boldsymbol{x})H\psi(\boldsymbol{x})/\psi^2(\boldsymbol{x}) = \psi^{-1}(\boldsymbol{x})H\psi(\boldsymbol{x})$. Since the local energy requires a division by $\psi(\boldsymbol{x})$ it is beneficial to rewrite the energy estimate in terms of the logarithm of the wavefunction, this removes the division and improves the numerical stability significantly. Plugging in the exponent of the logarithm of the wavefunction into the equation (6) and dividing by the wavefunction we get

$$E_L(\boldsymbol{x}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{3}\left[\frac{\partial^2 \log|\Psi(\boldsymbol{x})|}{\partial r_{ij}^2} + \left(\frac{\partial \log|\Psi(\boldsymbol{x})|}{\partial r_{ij}}\right)^2\right] + V(\boldsymbol{x}) \tag{10}$$

IV. Evolution strategies algorithm

Finding the set of parameters that minimizes the energy for a given system is a difficult problem. One of the main issues encountered when finding the gradient of the energy with respect to the parameters is the requirement of a large set of required samples of the wavefunction to compute it. Since this is a finite stochastic estimation of the energy its gradient does not actually represent the true gradient that minimizes the energy. As such utilizing stochastic gradient descent (SGD) methods is difficult. One way to avoid such a problem is to compute the energy estimate for a large set of different parametrizations of the wavefunction and find the optimal direction for optimization in such a way. One of the simplest methods of this type is the Evolution Strategies algorithm [10]. One of its benefits is its simplicity as well as not requiring computing the neural network gradient analytically, which improves the overall performance.

Evolution strategies is an iterative algorithm that given an initial parameter set guess θ_i samples a cost function around the guess $\theta_{i,j} = \theta_i + \Delta\theta_j$, where $\Delta\theta_j$ was sampled from a given random distribution. Then we take the average of the best samples and take that as the new best parameter set. This has the nice property of being resistant to strong fluctuations in the cost function.

While we can use the estimate of the energy as the cost function due to the lack of samples this leads to a rather poor estimate for the quality of the solution, to remedy this we can use a hybrid cost function which also includes the variation of the energy to avoid scenarios where the neural network produces sharp regions of high energy which are impossible to sample properly.

$$\mathcal{L} = E_{x\sim\psi_\theta^2}\left(\alpha E_L(x) + \beta\left(E_L(x) - E_{x\sim\psi_\theta^2}\left(E_L(x)\right)\right)^2\right), \tag{11}$$

where α and β are configurable parameters. These parameters should be chosen such that β is as small as possible while keeping the optimization stability. Since otherwise the network might converge to a mix of excited states because any solution to the Schrodinger equation has zero energy variation, not just the ground state one.

V. Results and implementation details

Computing large determinants required using a different representation of numbers to properly represent the non-normalized wavefunction values ranging from 10^{-30} to 10^{30}. To do so we stored the logarithm and the sign of the number instead of representing the values as floating points. Such a change greatly enhances the numerical stability of computing the wavefunction and allows us to directly use the more accurate expression for the local energy (10). To compute the spatial derivatives, we have used a second order finite difference stencil, which required $6N+1$ total evaluations of the wavefunction per local energy sample. We have also used $\alpha = 0.5$ and $\beta = 1.0$ as the parameters of the loss function.

A neural network with two hidden layers and eight neurons was used. Three determinants were used for all tests throughout. The number of parameters depends on the electron count but was around a thousand. The neural networks, the optimization algorithm and the Variational Monte Carlo setup have been written in HLSL for the DirectX API so that we can utilize the significant parallel computational capability of GPUs.

Firstly, we tested the accuracy of predicting the ground state energy of a select few first-row atoms, specifically Lithium, Carbon, Nitrogen and Neon in Table I. We compare the accuracy of the prediction to the exact value [5] as well as well-established methods like Hartree-Fock and Density Functional Theory. Even for such a small neural network this approach can find the energy more accurately than the other methods.

TABLE I. COMPUTED ENERGIES OF FIRST ROW ATOMS

Element	Exact, Ha	HF(CBS), Ha	DFT(LDA), Ha	Ours, Ha
Li	-7.47	-7.43	-7.33	-7.45
C	-37.84	-37.69	-37.46	-37.72
N	-54.58	-54.40	-54.12	-54.47
Ne	-128.94	-128.55	-128.23	-128.56

Comparing these results to Pfau et al. [5] we find out that our method achieves the specified level of accuracy significantly faster than FermiNet, by an order of magnitude. For instance, using our method, the calculation for Li requires approximately 2,537.5 trillion floating-point operations, while FermiNet necessitates 78,000 trillion operations.

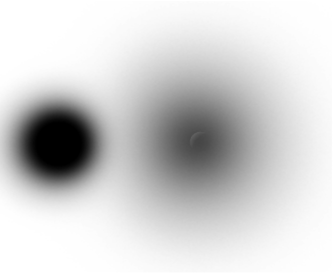


Fig. 1. The volumetric electron density of LiH after optimization.

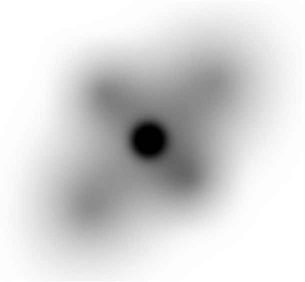


Fig. 2. The volumetric electron density of CH_4 after optimization.

We have also tested our predictions to experimental values for the first ionization energy for each of these atoms, the numerical values are in Table II. For the given setup we can get around 5% error for simple atoms, but we can clearly see that with an increasing number of electrons the accuracy of the prediction is reduced. This suggests that the size of the neural network must be increased with the size of the system to keep the accuracy constant.

TABLE II. FIRST IONIZATION ENERGY OF FIRST ROW ATOMS

Element	Ours, eV	Experiment, eV	Error, %
Li	5.14	5.40	4.76
C	10.91	11.26	3.09
N	14.46	14.53	0.50
Ne	18.44	21.56	14.45

For our last test we found the ground state of simple molecules such as lithium hydride (Fig. 1) and methane (Fig. 2). For LiH our method can very quickly converge to accuracies higher than HF, beating other VMC implementations in speed. In Fig. 3 we can see optimization process for LiH. The energy being the current Monte Carlo energy estimate, average energy is the exponential rolling average, with the true energy value provided by [5]. After 2 minutes the algorithm reached -8.033 Hartree, which is a 0.4% error compared to the exact value of -8.071. On methane we have reached a similar accuracy, -40.292 Hartree vs the exact value of -40.514 Hartree which is around 0.5% error, and the optimization process is shown on Fig. 4.

The Metropolis-Hastings algorithm allows us to visualize the electron probability distribution directly by building a histogram of the electrons, i.e., walker positions. We have plotted a high-resolution 3D volumetric histogram for our computed molecules, where the darker colors show higher electron probability for LiH and CH_4 shown on Fig. 1 and Fig. 2, the nuclei are shown as spheres. From these plots we can see that the electron probability is mainly concentrated around the nuclei with higher charge, where the core electrons reside.

We've also visualized the radial electron distribution histogram for a sole carbon atom, with the mean electron distance to the nuclei being 1.139 Bohr. The radius of the second peak is around 1.25 Bohr as can be seen in Fig. 5, which is consistent with the covalent radius of the atom of 1.341 Bohr.

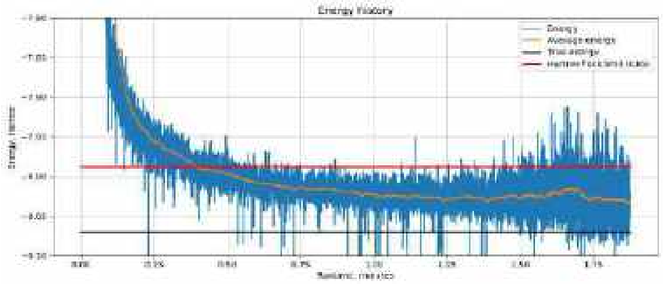


Fig. 3. Graph of system energy as a function of optimization time for LiH.

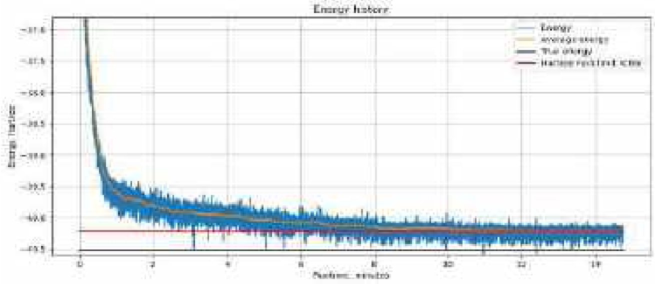


Fig. 4. Graph of system energy as a function of optimization time for methane molecule CH_4.

CONCLUSIONS

We have demonstrated that ground state wave functions can be efficiently identified using compact neural networks, with the assistance of evolution strategies (ES).

One of the standout features of ES is its notable resilience against being trapped in local minima. This characteristic, coupled with its stability, makes it advantageous when compared with more conventional gradient-based approaches. The employment of gradient-less techniques like ES comes with a distinct property of eliminating the need for computing the loss gradient in relation to parameters analytically, which in turn can increase the overall algorithmic performance.

Looking ahead, we recognize certain challenges related to the difficulty of scaling the size of the neural network while keeping its accuracy and speed. To address this, we plan to use a more refined variant of ES – the Limited Memory Matrix Adaptation Evolution Strategies [11]. This second-order optimization algorithm is well suited for Variational Monte Carlo (VMC) applications and presents a memory-efficient model, demanding only O(Nlog$N$) memory.

Building on this foundational work, we remain optimistic about the potential improvements we can make that may allow achieving near chemical accuracy under an ES algorithm. This can be realized by refining both the neural network architecture and the optimization methodology, all the while upholding the method's performance efficiency. This method can be extended to study the properties of solid bodies, by moving into second quantization [12].

The generalizability of neural networks means that we can describe any kind of physical property of any system if we increase the neural network size sufficiently. This allows for studying the characteristics of new materials which were too large to be tackled by classic CCSD(t) or FCI methods, providing a new tool to efficiently find novel materials.

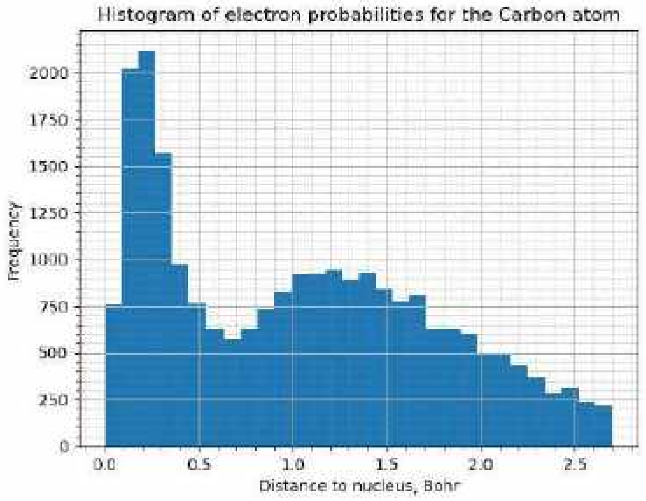


Fig. 5. Histogram of electron probabilities around a Carbon nucleus after optimization.

Such precision is essential not just for understanding how molecules change and interact, but also for exploring the electronic behaviors of materials, like superconductors.

REFERENCES

[1] J. Behler and M. Parrinello, "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces," Phys. Rev. Lett., vol. 98, no. 14, pp. 146401, Apr. 2007. doi: 10.1103/PhysRevLett.98.146401

[2] M. Moroz. Predicting Equilibrium Geometries of Large Multicomponent Systems with Neural Networks / M. Moroz, O. Bovgyra, V. Franiv, V. Dzikovskyi // Proceedings of the Xth International Scientific and Practical Conference "Electronics and Information Technologies" (ELIT-2018), Lviv-Karpaty village, August 30 – September 2 2018. – Lviv: Ivan Franko National University of Lviv, 2018. – P. A61-A64. doi.org/10.30970/elit2018.A19.

[3] A. Szabó and N.S. Ostlund, "Modern quantum chemistry: Introduction to advanced electronic structure theory," Proceedings, 1982.

[4] J. Hermann, J. Spencer, K. Choo, A. Mezzacapo, W. M. C. Foulkes, D. Pfau, G. Carleo, and F. Noé, "Ab initio quantum chemistry with neural-network wavefunctions," Nature Reviews Chemistry, 2023. doi: 10.1038/s41570-023-00516-8

[5] D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes, "Ab initio solution of the many-electron Schrödinger equation with deep neural networks," Phys. Rev. Res., vol. 2, no. 3, pp. 033429, Sep. 2020.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, Oct. 1, 1986. DOI: 10.1038/323533a0

[7] T. Kato, "On the Eigenfunctions of Many-Particle Systems in Quantum Mechanics," Proceedings, 2011.

[8] W. L. McMillan, "Ground State of Liquid He4," Phys. Rev., vol. 138, no. 2A, pp. A442–A451, Apr. 1965. doi: 10.1103/PhysRev.138.A442

[9] W. K. Hastings, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," Biometrika, vol. 57, no. 1, pp. 97–109, 1970.

[10] T. Salimans, J. Ho, X. Chen, and I. Sutskever, "Evolution Strategies as a Scalable Alternative to Reinforcement Learning," 2017.

[11] I. Loshchilov, T. Glasmachers, and H.-G. Beyer, "Limited-Memory Matrix Adaptation for Large Scale Black-box Optimization," IEEE Transactions on Evolutionary Computation, vol. PP, 2017. DOI: 10.1109/TEVC.2018.2855049

[12] N. Yoshioka, W. Mizukami, and F. Nori, "Solving quasiparticle band spectra of real solids using neural-network quantum states," Communications Physics, vol. 4, no. 1, pp. 106, May 21, 2021. DOI: 10.1038/s42005-021-00609-0

The Influence of the Input Array on the Learning Error of a Multilayer Neural Network

Serhiy Sveleba
Department of Optoelectronics and Information Technologies
Ivan Franko National University of Lviv,
Lviv, Ukraine
serhiy.sveleba@lnu.edu.ua
ORCID:0000-0002-0823-910X

Ivan Katerynchuk
Department of Optoelectronics and Information Technologies
Ivan Franko National University of Lviv,
Lviv, Ukraine
ivan.katerynchuk@lnu.edu.ua
ORCID: 0000-0001-8877-8324

Ivan Kunyo
Department of Optoelectronics and Information Technologies
Ivan Franko National University of Lviv,
Lviv, Ukraine
ivan.kuno@lnu.edu.ua
ORCID: 0000-0001-6092-7949

Natalia Sveleba
Department of Optoelectronics and Information Technologies
Ivan Franko National University of Lviv,
Lviv, Ukraine
incomlviv@gmail.com

Ostap Semotiuk
Department of Computer Printing Engineering
Ukrainian Academy of Printing of Lviv,
Lviv, Ukraine
incomlviv@gmail.com

Volodymyr Kotsun
Lviv Branch of Private Higher Education Establishment "European University"
Lviv, Ukraine
incomlviv@gmail.com

Abstract — **For a multilayer neural network (MLNN) with three hidden layers, the influence of the input array size on the learning process was studied when recognizing printed digits. A study was conducted both with and without the use of optimization training methods such as AMSGrad, and AdaDelta. The learning error of this neural network was analyzed using the Fourier spectra of the error function and constructing branching diagrams when analyzing the logistic function that describes the doubling of existing local minima. The Fourier spectra were found to indicate a more significant number of harmonics in the training process for arrays with smaller dimensions. Increasing the input array size for representing the digits (3x5, 4x7, 28x28 pixels) leads to more homogeneous learning process.**

It is demonstrated that the existence of error function harmonics throughout the range of learning parameter variations, given a fixed learning rate, is associated with the heterogeneity of the input array. In other words, the heterogeneity of the input array provokes the emergence of local minima and, consequently, the more complex behavior of the error function for the learning parameters. Therefore, the heterogeneity of the input array, under the condition of a fixed learning rate, acts as a catalyst for the neural network retraining process.

Keywords* — *Multilayer neural network; AMSGrad and AdaDelta optimization methods; Python.

I. Introduction

The selection of data set for MLNN training is one of the crucial stages. The training data set should satisfy several criteria, namely:

- The data should accurately represent real-world situations in the subject area.
- Contradictory data in the training sample can lead to poor training quality of the network.
- Typically, the number of records in the sample should exceed the number of connections between the neurons in the network by several orders of magnitude.

Another requirement is the sufficient representativeness of the training data set: the better the network generalizes the training, the denser and more uniformly distributed the training data should be in the input space. The network can perform proper generalization by interpolating the input data if the test data is always provided between closely located training patterns.

When considering neural networks, the role of input data selection in the training process has been emphasized [1, 2], specifically its influence on the training error. For instance, in [3], it was noted that increasing the size of the sample, both in terms of the array of the digit itself and the sample of digits, led to improved training and digit recognition accuracy. It was mentioned that although the size of the input data array increases, it allows for a reduction in the number of iterations. As a result, the recognition process for all digits occurs with minimal error.

Nevertheless, the question of the impact of the input array's homogeneity on a neural network's training process arises. Specifically, this article examines the influence of the dimensionality of the input array representing the digit, i.e., representing the digit as a two-dimensional array with varying numbers of pixels. Now let us consider the effect of a 3x5, 4x7, and 28x28 pixel array of digits on the learning process of a multilayer neural network both with and without the AMSGrad and Adadelta optimization methods.

II. Methodology

In the Python programming environment, a program was written to describe a multilayer neural network with hidden layers for recognizing printed digits. The array for each digit consisted of a set of "0" and "1" in 3x5, 4x7, and 28x28 sizes. For example, in the considered array of 4x7 pixels, each digit (for example, the digit "5") contained four variants of possible digit distortions (Num52; Num53; Num54; Num55). In other words, the digit "5" was represented by the following array:

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

Num51=[1,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,1]
Num52=[1,1,1,0,1,0,0,0,1,0,0,0,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,1]
Num53=[0,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,1]
Num54=[1,1,1,1,1,0,0,0,0,0,0,0,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,1]
Num55=[1,1,1,1,1,0,0,0,1,0,0,0,1,0,1,1,0,0,0,1,0,0,0,1,1,1,1,1]

The neural network under consideration contained 3 hidden layers with 28 neurons in each layer. The selection of the number of hidden layers and neurons in each of them was determined by the slightest learning error for digit recognition. According to [4], this three-layer neural network had 15, 28, and 784 neurons in each layer, respectively, for input arrays of 3x5, 4x7, and 28x28 pixels.

The following logistic function was used to analyze the error function:

$$x_{n+1} = a - x_n - x_n^2,$$

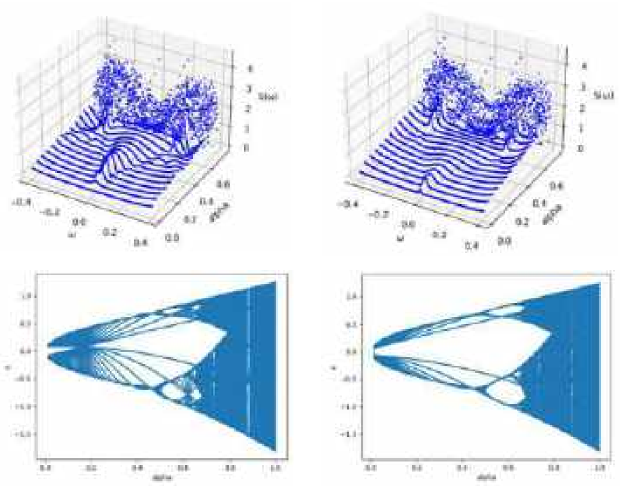
where n is the number of iterations, and a is a parameter. In the case of the AdaDelta optimization training method, a represents the hyperparameter *rho* (a = *rho*), typically set to 0.9. It determines the contribution of the squared gradient of the objective function. In the case of the AMSGrad optimization training method or when no optimization training methods are used, a represents the stationary learning rate alpha (a = *alpha*).

The choice of this logistic mapping is motivated by its description of the process of doubling the frequency of oscillations [5]. In our case, this process is caused by the emergence of local minima when approaching the global minimum.

III. No Optimization Methods

The influence of data retrieval on the learning process of MLNN at no optimization learning methods. It is known [3] that one of the parameters that significantly influences the training error of a neural network in the absence of optimization methods is the selection of input data samples. Specifically, in our case, it is the number of pixels allocated for representing a single digit. Fig. 1 shows the Fourier spectra and branching diagram for two arrays representing of printed digits. The first one is a 3x5 pixel array (Fig. 1a), and the second one is a 4x7 pixel array (Fig. 1b). Comparing the Fourier spectra of these two arrays, it should be noted that for the representation of a digit in the 3x5 format, the Fourier spectrum exhibits a greater number of existing harmonic functions of the error. It indicates a heterogeneous neural network learning process, as indicated by the branching diagrams. The branching diagram for the 3x5 digit representation format demonstrates a more significant number of local minimum duplications. According to the branching diagram shown in Fig. 1a, even at the initial stage of neural network training, the learning error function for each neuron is a complex functional dependency. Increasing the number of pixels in the digit representation leads to a reduction in this heterogeneous behavior of the error function (branching diagram in Fig. 1b). That is also demonstrated by the Fourier spectra in Fig. 1b. With a learning rate *alpha*>0.5, the process of local minimum appearance becomes noticeable.

Further increasing the learning rate, the logistic error function begins to describe a process of doubling the number of local minima in the error function. In the end, this leads to the emergence of chaotic behavior in the error function and, consequently, in the neural network. Regardless of the pixel array used for digit representation, chaotic behavior practically occurs at the same learning rate values (*alpha* > 0.7). As it is known, the appearance of local minima in the error function for the number of iterations or the learning rate is caused by the retraining of individual neurons. Increasing the number of neurons involved in the retraining process leads the neural network to transition into a chaotic mode (repeated passing of the global minimum). The absence of the neural network learning process characterizes this mode.



a) b)

Fig. 1. Fourier spectra vs learning rate and branching diagram for the digit "0" when the digit is set to an array of a) 3x5 and b) 4x7 pixels at 100 iterations

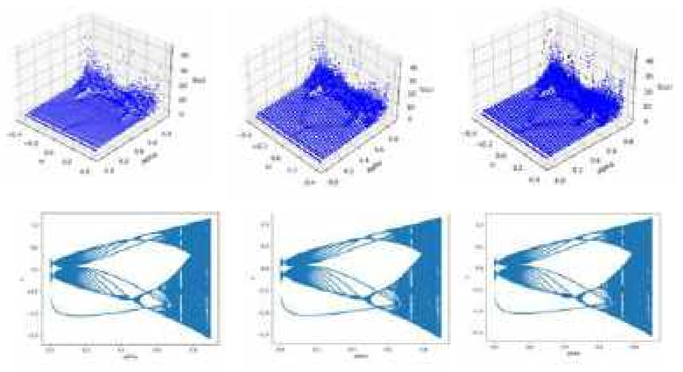
IV. AMSGrad

The influence of data retrieval on the learning process of a multilayer neural network when applying the amsgrad optimization learning method.

Now let us consider the influence of the dataset when applying one of the training optimization methods. The AMSGrad method was chosen as the optimization method, effectively preventing the retraining of the neural network [6]. For comparison, Fig. 2 shows the branching and Fourier spectra diagrams for two arrays of 3x5 pixels (Fig. 2a) and 4x7 pixels (Figure 2b).

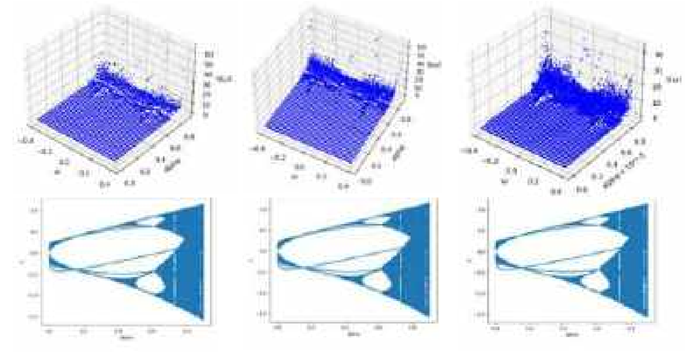
The branching diagrams indicate the presence of a greater number of existing harmonics for the 3x5 pixel array. These periodicities exist throughout the learning rate range, suggesting a correlation with the heterogeneity of the input array. According to the branching diagram, these periodicities associated with the heterogeneity of the input array contribute to the emergence of local minima and, therefore, the more complex behavior of the error function for the learning parameters. Hence, heterogeneities in the input array act as catalysts for the retraining process of the neural network.

Comparing the Fourier spectra diagrams of the given arrays, it should be noted that for the 3x5 pixel array, the Fourier spectra contain more harmonics. They are characterized by oscillations that exist even before the system enters a chaotic state. As for the doubling process of the existing oscillations' frequency, it follows a more complicated scenario with the emergence of additional ones.



digit =0 digit =1 digit =2

Fig. 2 a. Fourier spectra versus learning rate and branching diagram for the digits "0", "1", and "2" when the digit is set to an array of 3x5 pixels at 100 iterations. Used AMSGard method



digit =0 digit =1 digit =2

Fig. 2 b. Fourier spectra versus learning rate and branching diagram for the digits "0", "1", and "2" when the digit is set to an array of 4x7 pixels at 100 iterations. Used AMSGard method

V. Adadelta

The influence of data retrieval on the learning process of a multilayer neural network when applying the adadelta optimization learning method.

Adadelta is a further extension of RMSProp, developed to improve the algorithm's convergence and eliminate the need to specify the initial learning rate [7] manually. The idea is to improve on two major drawbacks of the AdaGrad method:

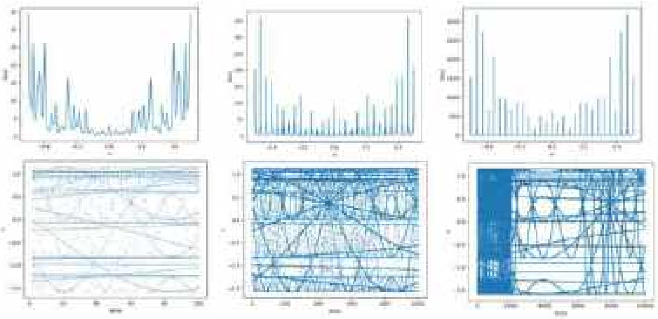
1) Continuous decrease of the learning rate during training.

2) The need to manually choose a global learning rate.

Figure 3a shows the Fourier spectra of the error function for the digit "0" after the last hidden layer in a three-layer neural network. The Fourier spectra were calculated for 28 neurons. Since the digit array was 4x7 pixels, the learning rate is automatically selected according to the Adadelta algorithm for each neuron in the array. Therefore, the Fourier spectra at 100 iterations exhibit oscillations of varying amplitudes. Also, it indicates the involvement of all neurons in the learning process. The obtained spectra represent a "differentiable" function, indicating the absence of "higher" harmonics.

These Fourier spectra were obtained with a hyperparameter value of rho=0.9. This value is determined by the optimization algorithm of this method by default [7]. The branching diagram shown in Figure 3b indicates that the training of neurons is not homogeneous. Specifically, the diagram reveals neurons for which the learning error function depends on the number of iterations, although their percentage is small. Increasing the number of iterations (up to 1000) leads to the appearance of higher harmonics in the Fourier spectrum of the learning error function for each neuron.

Additionally, an increase in the power of the main harmonics associated with each neuron can be observed. The resulting branching diagram is characterized by the emergence of a chaotic state with the appearance of attractors corresponding to the fundamental periodicities of the system. It is also worth noting that about one-third of the neurons exhibit a monotonically slowly changing behavior of the error function for the number of iterations.

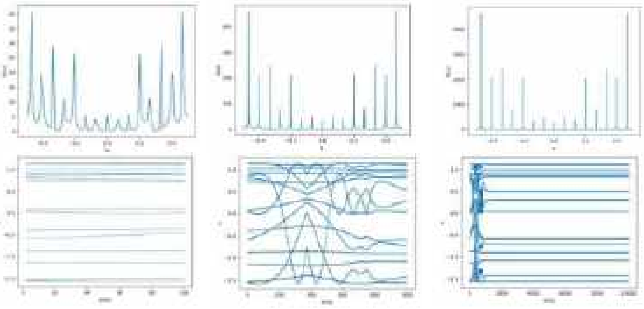


a) iterations =100 *b)* iterations =1000 *c)* iterations =10000

Fig. 3. Fourier spectra and branching diagrams for a three-layer neural network with 28 neurons in each layer, an input array of 4x7 pixels, and rho=0.9 for the digit "0" using the AdaDelta optimized learning method

Further increasing the number of iterations (up to 10000) causes the disappearance of higher harmonics and increases the power of the main ones (Figure 3a). In this case, the bifurcation diagram is characterized by the disappearance of chaotic behavior and the emergence of a monotonic, nearly constant behavior of the error function on each neuron. In other words, the Adadelta optimization learning method achieves an optimal learning rate selection for each neuron, which starts working correctly after 2000 iterations.

If we analyze the learning process with an input array of 3x5 pixels compared to an array of 4x7 pixels, we should note a monotonic, not chaotic neural network training process. At 100 iterations, the error function on each neuron changes linearly and monotonically. Increasing the number of iterations to 1000 provokes a non-linear behavior of the error function on one-fourth of the neurons. Further increasing of iterations up to 10000 causes a return to linear and monotonic behavior of the error function on each neuron. Therefore, when reducing the size of the input array, the learning process, with the application of the Adadelta optimization learning method that provides automatic learning rate adjustment, proceeds without the occurrence of chaotic behavior, although it is accompanied by retraining of several neurons (Fig. 4 at 1000 iterations).



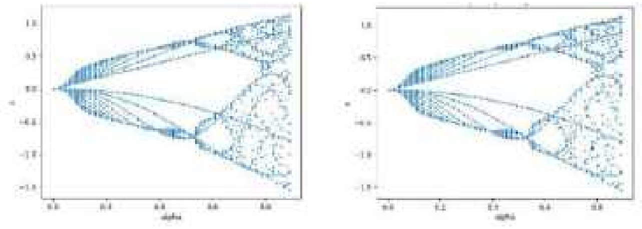
a) iterations =100 b) iterations =1000 c) iterations =10000

Fig. 4. Fourier spectra and branching diagrams for a three-layer neural network with 15 neurons in each layer, an input array of 3x5 pixels, and rho=0.9 for the digit "0" using the AdaDelta optimized learning method

VI. 28X28 PIXELS SET

The process of MLNN training using AMSGrad optimization learning methods when sampling an input array of 28x28 pixels from a set of handwritten digits.

In Fig. 5, branching diagrams are shown for the handwritten digit "0", which was defined by a 28x28 pixel array in 600 (Fig. 5a) and 300 variations (Fig. 5b). The obtained branching diagram is similar to the one obtained when using a 3x5 and 4x7 pixel array to represent the digit. The difference in the case of a 28x28 sample is observed in the learning rate value at which the process of doubling the number of existing minima starts to occur. For the 28x28 pixel sample, this value is approximately 0.55, while for the 3x5 and 4x7 pixel samples, it is approximately 0.5.



a) 600 variations *b*) 300 variations

Fig. 5. Branching diagram for the digit "0", for different variations of the digits "0": a) 600 variations and b) 300 variations, at 10 iterations using the AMSGrad optimization learning method.

It should be noted that the obtained learning rate values at which the process of doubling the number of local minima (the emergence of a retraining process) occurs are independent of the number of representations of the given digit. In our opinion, the increase in the learning rate value at which the retraining process becomes noticeable is associated with the fact that increasing the retrieval size leads to greater homogeneity of the input array compared to ones with smaller sizes.

CONCLUSIONS

In this paper, the authors have studied the influence of the input array size on the learning process for a multilayer neural network, both with and without applying AMSGrad and Adadelta optimization methods. The Fourier spectra of the error function and branching diagrams of the logistic error function in the coordinates number of iterations and alpha (with the AdaDelta optimization learning method, the number of iterations, and with the AMSGrad optimization learning method and in the absence of optimization learning methods, alpha) show that the learning process is sensitive to the size of the input array. Increasing the array of the digit itself leads to a homogeneous learning process. It has been found that Fourier spectra indicate the presence of a greater number of harmonics during the training process for arrays with smaller dimensions.

It is shown that the existence of error function harmonics throughout the range of parameter variations in the training process, assuming a fixed learning rate, is associated with the heterogeneity of the input array. In other words, the heterogeneity of the input array provokes the process of local minima and hence a more complex behavior of the error function as a function of the learning parameters. Thus, the heterogeneity of the input array, under the condition of a fixed learning rate, acts as a catalyst for overfitting in the neural network. The opposite principle is observed when using automatic learning rate adjustment (e.g., with the AdaDelta optimized training method). Specifically, reducing the input array size leads to a homogeneous neuronal training process. However, this process is accompanied by an increase in the number of iterations to establish a monotonic and homogeneous learning process with increasing heterogeneity of the input array (for example, learning different digits).

REFERENCES

[1] Subbotin, S. O. (2020). Neironni merezhi: teoriia ta praktyka: navchalnyi posibnyk [Neural Networks: Theory and Practice. A textbook]. Zhytomyr, O.O. Yevenok Publishing House, 184 p. (in Ukrainian).

[2] Savchuk, T. O., Yarema, Ye. O. (2005). Vykorystannia neironnykh merezh dlia rozpiznavannia symvoliv [The use of neural networks for characters recognition]. Scientific and Technical Journal "Data Recording, Storage & Processing," Institute of Information Recording Problems of the National Academy of Sciences of Ukraine, Kyiv, Vol. 7, No. 4, p.78-84, ISSN 1560-9189. (in Ukrainian).

[3] S. Sveleba, I. Katerynchuk, I. Kuno, Y. Shmyhelskyy, N. Sveleba, "The Role of Sample Size in Multilayer Neural Networks" Proceedings of the 16th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, TCSET 2022. – 22-26 February 2022. – Lviv-Slavske, Ukraine. – P. 767-770 DOI: 10.1109/TCSET55632.2022.9767025

[4] S. Sveleba, I. Katerynchuk, I. Kunyo, I. Karpa, O. Semotyuk, Ya. Shmygelsky, N. Sveleba, V. Kunyo "Chaotic States of a Multilayer Neural Network", Electronics and information technologies. Is. 13. pp. 20–35, 2021. DOI: http://dx.doi.org/10.30970/eli.16.3

[5] Yu. Taranenko "Information entropy of chaos". [Online]. Available: https://habr.com/ru/post/447874/.

[6] Sebastian Ruder, Aylien Ltd., Dublin "An overview of gradient descent optimization algorithms□". [Online]. Available: https://doi.org/10.48550/arXiv.1609.04747

[7] Matthew D. Zeiler. "Adadelta: an adaptive learning rate method". [Online]. Available: https://doi.org/10.48550/arXiv.1212.5701

The Influence of Sampling Parameters on the Learning Error of a Multilayer Neural Network

Serhiy Sveleba
Department of Optoelectronics and Information Technologies
Ivan Franko National University of Lviv,
Lviv, Ukraine
serhiy.sveleba@lnu.edu.ua
ORCID:0000-0002-0823-910X

Ivan Katerynchuk
Department of Optoelectronics and Information Technologies
Ivan Franko National University of Lviv,
Lviv, Ukraine
ivan.katerynchuk@lnu.edu.ua
ORCID: 0000-0001-8877-8324

Ivan Kunyo
Department of Optoelectronics and Information Technologies
Ivan Franko National University of Lviv,
Lviv, Ukraine
ivan.kuno@lnu.edu.ua
ORCID: 0000-0001-6092-7949

Natalia Sveleba
Department of Optoelectronics and Information Technologies
Ivan Franko National University of Lviv,
Lviv, Ukraine
incomlviv@gmail.com

Serhiy Velhosh
Department of Radiophysics and Computer Technologies
Ivan Franko National University of Lviv,
Lviv, Ukraine
serhiy.velhosh@lnu.edu.ua
ORCID: 0000-0002-0011-6359

Volodymyr Kotsun
Lviv Branch of Private Higher Education Establishment “European University”
Lviv, Ukraine
incomlviv@gmail.com

***Abstract* — The influence of sampling parameters on the learning process for a multi-layer neural network (MLNN) when recognizing printed numbers was studied. MLNN had three hidden layers of 28 neurons in each. Adam, AdamMax, and AMSGrad optimization learning methods were used. Testing the learning error of this neural network was carried out by constructing maps of the dynamic modes of the error function in alpha – beta2 coordinates. Alpha is the stationary value of the learning step, beta2 is the optimization parameter of the error function, which determines the contribution of the square of the gradient of the error function. It was established that the maps of dynamic regimes testify to the existence of a greater number of observed periodicities for input arrays with smaller dimensions. An increase in the set of the number itself and the variants of their distortions leads to a decrease in the learning error. The heterogeneity of the input array contributes to the process of the appearance of local minima, and therefore to more complex behavior of the error function from the learning parameters. The obtained results prove that the heterogeneity of the input array is a catalyst for the retraining process of the neural network.**

Keywords — Multilayer neural network; Adam, AdamMax, i AMSGrad optimization methods; Python.

I. Introduction

When considering neural networks, the influence of the sampling of input data on the learning process is often emphasized [1, 2], namely the influence on the accuracy of learning. For printed numbers, it was noted in the work [3] that increasing the size of the sample that specifies the number and the sample that includes the distortion of this number led to an increase in the accuracy of learning and speeding up the process of recognizing numbers. In particular, it was noted in [3] that although the array of input data is growing, it allows to reduce the number of iterations. At the same time, the process of recognizing all digits takes place with minimal error.

II. Methodology

A program was written in the Python programming environment that describes a MLNN with hidden layers for recognizing printed digits. The array of each number consisted of a set of "0" and "1" of size 3x5 or 4x7. For example, in the 4x7 pixel array we considered, each number (for example, the number "5", Num51) still contained four variants of possible distortions of the number (Num52; Num53; Num54; Num55), and three variants that did not correspond to any number (Numt1; Numt2 ; Numt3). That is, the number "5" was specified by the following array:

```
Num51=[1,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,1]
           1     1     1     1
           1
           1
           1     1     1     1
                             1
                             1
           1     1     1     1
Num52=[1,1,1,0,1,0,0,0,1,0,0,0,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,1]
Num53=[0,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,1]
Num54=[1,1,1,1,1,0,0,0,0,0,0,0,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,1]
Num55=[1,1,1,1,1,0,0,0,1,0,0,0,1,0,1,1,0,0,0,1,0,0,0,1,1,1,1,1]
Numt1=[0,0,0,0,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
Numt2=[1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1]
Numt3=[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
```

The considered MLNN contained 3 hidden layers with 28 neurons in each layer. The choice of the number of hidden layers and neurons in each of them was determined by the smallest learning error for number recognition. According to [3], this is a three-layer neural network with 28 neurons in each layer. Maps of dynamic regimes provide a fairly complete and visual representation of the behavior of a dynamic system. A map of dynamic modes is a diagram on a plane where two parameters are plotted along the coordinate axes and the boundaries of the regions of different dynamic modes are shown. Constructed maps of dynamic regimes are correlated with other ways of presenting this information, such as a map of Lyapunov or Arnold indicators. Both of these methods most often duplicate the information obtained with the help of maps of dynamic regimes.

Two-dimensional mappings, as well as one-dimensional ones, are given by recurrence relations of the form:

$$x_{n+1}=f(x_n, y_n);$$

$$y_{n+1}=g(x_n, y_n).$$

In two-dimensional mappings, you have to deal with points on the plane, that is, with several numbers that will specify the coordinates of the points. Two-dimensional

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

mappings come into consideration in various ways. Some are the result of the generalization of one-dimensional representations, others model some phenomenon characterized by discrete time. Often, two-dimensional mappings arise as difference schemes during the numerical solution of systems of differential equations. In our case, the role of the function $f(x_n, y_n)$ is the error – the difference between the expected value and the real value. The two-dimensional mapping constructed in this way depends on the values of the parameters $a = alpha$ and $b = beta2$ (where *alpha* is the stationary value of the learning step, *beta2* is the error function optimization parameter, and determines the contribution of the square of the gradient of the error function). The choice of these *alpha* and *beta2* parameters as parameters a and b is due to the sensitivity of the learning process to these parameters. As in the case of one-dimensional mappings, two-dimensional mappings allow the use of bifurcation diagrams. However, since the number of parameters, as a rule, is greater than one, instead of bifurcation diagrams, maps of dynamic modes look better.

III. 3X5 DATA SET

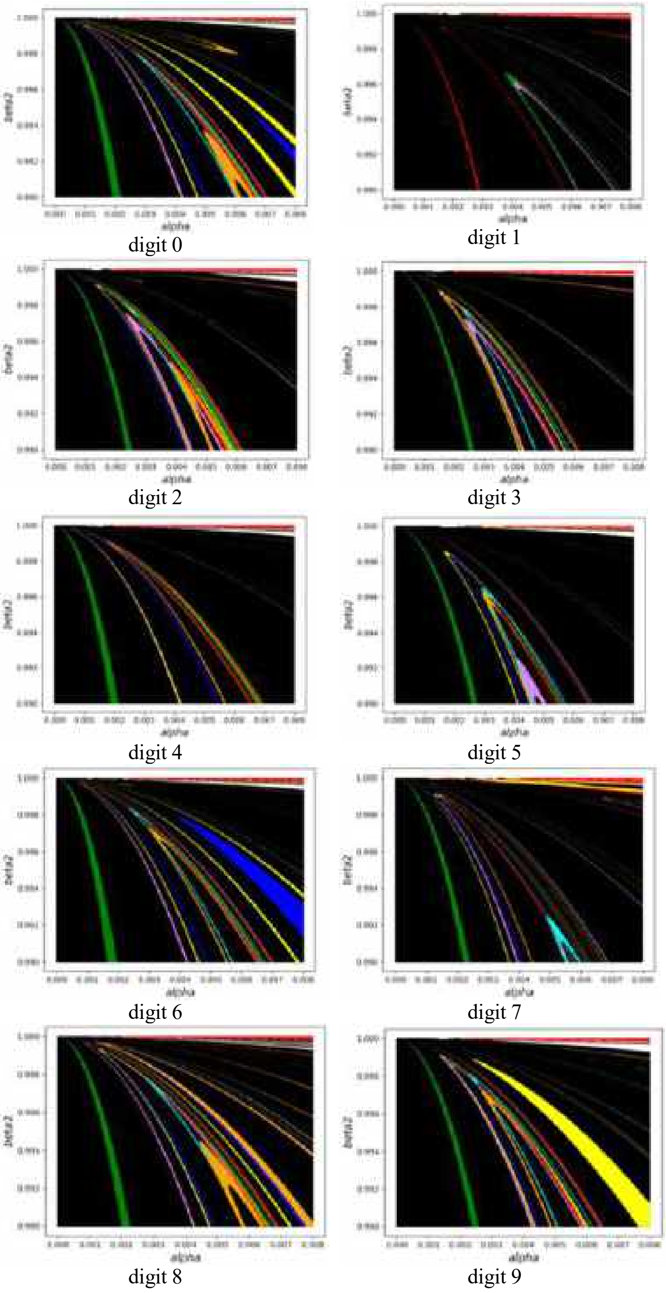
Let consider the maps of dynamic modes at small values of the number of iterations (5 and 10) for optimization methods of neural network training, such as Adam, AdaMax, AMSGrad. The choice of these optimization methods is due to the fact that they are in demand and give good results. We will start the consideration of the maps of dynamic modes with a small number of iterations in order to trace the full picture of the dynamics of neural network learning.

Fig. 1 and Fig. 2 show maps of dynamic modes for printed numbers at 5 and 10 iterations, respectively. According to Fig. 1, all figures are characterized by their own map of dynamic modes. Maps of dynamic modes corresponding to the numbers "0" are characterized by a larger palette of colors; "2"; "3"; "5"; "6"; "9". The genesis of all observed periodicities begins to be traced around alpha = 0.002 and beta2 = 0.999. With an increase in the alpha learning rate and a decrease in the beta2 optimization parameter, there is some expansion of the interval of existence of the observed periodicities and their divergence. Highlighting the existing periodicity that occupies most of the area on the map of dynamic modes is problematic, since the observed periodicities are approximately the same size.

Using a color palette, we can easily determine the type and period of the regime that exists in the system in a certain range of parameter changes. The colors determine the period of the corresponding periodic movement: red (red)-1; orange (orange)-2; yellow (yellow)-3; green (green)-4; blue (cyan)-5; blue (blue) -6; violet (violet)-7; black (black) – all others.

Analyzing the observed maps of dynamic modes, for different numbers, it can be stated that, given an array of numbers in the size of 3x5 pixels, the map of dynamic modes is characterized by a wide range of periodic movements. Increasing the number of iterations to 10 slightly modifies the map of dynamic modes. In particular, there is an increase in the interval of existence of periodic oscillations with a lower periodicity (Fig. 2).

This is quite evident for the numbers "0"; "2"; "4"; "6"; "8". When the *beta2* parameter increases with constant step learning, for all digits, the interval of existence of periodicities and their number is decreases. As the dynamic mode maps show, the detected periodicities are sensitive to the number of iterations. At the same time, an increase in the share of periodicities with the smallest period can be observed as the number of iterations increases (Fig. 2).



digit 0

digit 1

digit 2

digit 3

digit 4

digit 5

digit 6

digit 7

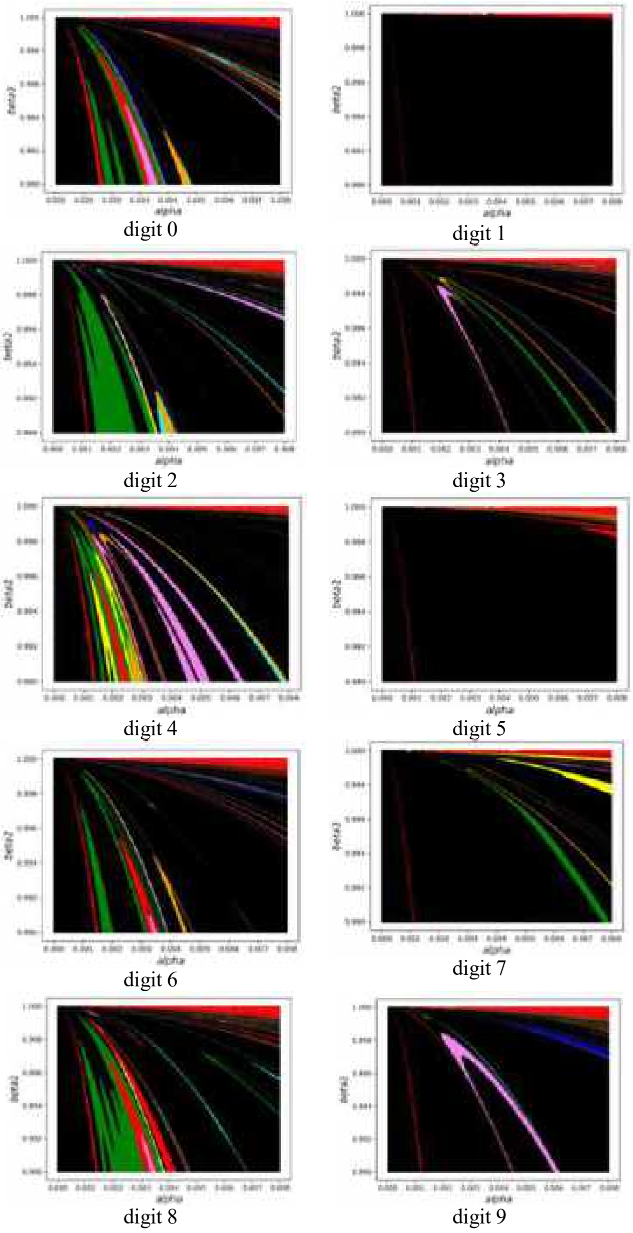
digit 8

digit 9

Fig. 1. Maps of the dynamic learning modes of a three-layer neural network for printed digits are given by an array of 3x5 pixels, subject to the use of the AdaMax optimization method, the number of iterations is 5, the optimization parameter *beta2*=0.9

Therefore, the periodicities that describe the behavior of the error function converge when the optimization parameter beta2 increases. Similar results were obtained for the optimization methods of learning Adam and AMSGrad.

Therefore, the obtained maps of dynamic modes when applying optimization methods of learning Adam, AdaMax, AMSGrad are practically the same. That is, with this sample of numbers representation (3x5), the learning process takes place according to the same scenario, and changing the optimization method of learning practically does not affect the existing periodicities. Insignificant differences in the behavior of the existing periodicities can be traced only around beta2 ≈ 0.999, at the values of the learning speed, which corresponds to a satisfactory learning process.



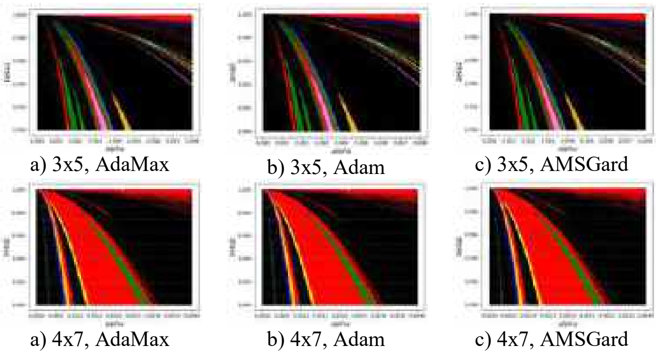
digit 0 digit 1
digit 2 digit 3
digit 4 digit 5
digit 6 digit 7
digit 8 digit 9

Fig. 2. Maps of the dynamic learning modes of a three-layer neural network for printed digits are given by an array of 3x5 pixels, subject to the use of the AdaMax optimization method, the number of iterations is 10, the optimization parameter *beta2*=0.9

IV. 4X7 DATA SET

In Fig. 3 maps of dynamic modes are given for the digits "0" when using the optimization learning methods AdaMax (Fig. 3, a), Adam (Fig. 3, b) and AMSGrad (Fig. 3, c). The obtained maps of dynamic modes are almost similar for the considered figure. These maps of dynamic regimes are characterized by the entire palette of existing periodicities, and the periodicities with the smallest periods are dominant. The difference between the maps of dynamic modes for different numbers is manifested only in the peculiarities of the behavior of periodicities with higher periods. An increase in the beta2 parameter at a constant learning rate for all digits causes a decrease in the interval of existence of different periodicities and their number. The detected periodicities with the array of 4x7 pixels, as well as with the array of 3x5 pixels, are sensitive to the number of iterations. At the same time, an increase in the share of periodicities with the smallest period can be observed as the number of iterations increases. Therefore, the periodicities that describe the behavior of the error function in the case of a 4x7 pixel number display array, as well as in the case of a 3x5 array, converge when the *beta2* optimization parameter increases.

Comparison of maps of dynamic modes with 3x5 and 4x7 arrays, under the same training conditions (the range of the *beta2* optimization parameter change, the training step) testify to a different palette of existing periodicities. Namely, for the numbers given by the 3x5 pixel array, a lager palette of existing periodicities is observed. The periodicities with the smallest period (red (red)-1; orange (orange)-2; yellow (yellow)-3; green (green)-4) become dominant on the maps of dynamic modes for the numbers specified by the 4x7 array. An increase in the interval of existence of periodicities with the smallest period, as well as an increase in the number of such areas on the map of dynamic modes with an increase in the number representation array, testifies to an increase in the uniformity of the learning process. That is, an increase in the number presentation array leads to a monotonous and homogeneous process of learning neurons, reducing the retraining of neurons. Thus, it can be argued that the size of the number representation array has a significant impact on the learning process. Based on the obtained results, it can be assumed that the learning process is influenced by the heterogeneity of the input array.



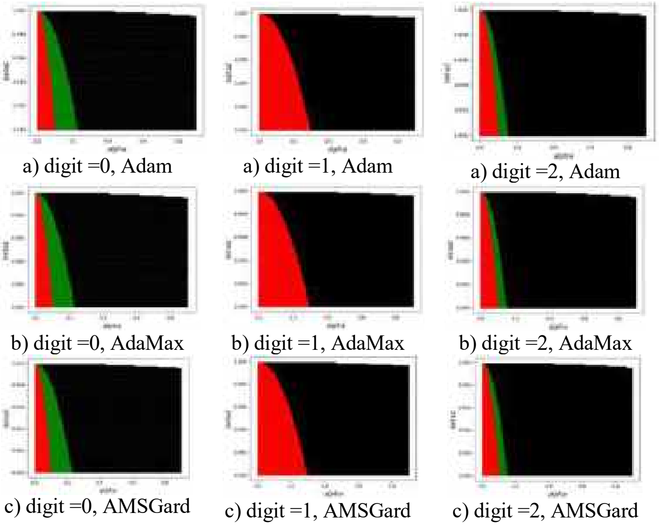
a) 3x5, AdaMax b) 3x5, Adam c) 3x5, AMSGard
a) 4x7, AdaMax b) 4x7, Adam c) 4x7, AMSGard

Fig. 3. Map of dynamic modes in the alpha and beta2 axes when applying the optimization learning method a) AdaMax, b) Adam, c) AMSGrad, provided the number of iterations is 10, the optimization parameter *beta2*=0.9

V. HOMOGENOUS DATA SET

Reducing the heterogeneity of the input array can be done by excluding from the array each digit, the digit that is the most distorted compared to the others. In the array we considered, each digit contained four variants of possible distortions of the digit (Num52; Num53; Num54; Num55), and three variants that did not correspond to any digit (Numt1; Numt2; Numt3). The assignment of such an array of numbers was justified, according to work [3], by a lower learning error for recognition of printed numbers by a three-layer neural network. In our case, according to the conducted studies of the maps of dynamic modes, the appearance of periodicities with long periods in the case of a 3x5 pixel number display array is possibly connected with the existence of a variant in the sample that does not correspond to any of the numbers. Therefore, consider an array of sample numbers that does not include such an option.

Fig. 4 shows maps of dynamic modes at values of the learning step $0.0001 < alpha < 0.9$. The obtained maps of dynamic modes for different optimization methods are identical and are defined by three colors. Red – a period equal to one, green – a period equal to 4, and black – a period greater than 7. The maps of dynamic modes shown in Fig. 4 do not depend on the considered optimization methods of training. The difference between the maps of the dynamic modes can be seen only when moving from one figure to another. The transition to a state that can be characterized as chaotic (black color) is carried out by doubling the number of periodicities. It is about such a transition to chaos, i.e. due to the doubling of the quantities of existing periodicities, that the branching diagrams indicate.



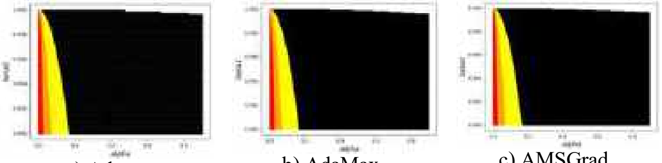
a) digit =0, Adam a) digit =1, Adam a) digit =2, Adam

b) digit =0, AdaMax b) digit =1, AdaMax b) digit =2, AdaMax

c) digit =0, AMSGard c) digit =1, AMSGard c) digit =2, AMSGard

Fig. 4. Maps of the dynamic learning modes of a three-layer neural network for printed digits given by an array of 4x7 pixels, for the optimization methods of learning a) Adam, b) AdaMax, c) AMSGrad, provided the number of iterations is 10, the optimization parameter *beta2*=0.9.

The difference between the learning processes for different digits is due to the different intervals of existence of periodicities corresponding to periodicities with a longer period. This indicates that each number has its own learning "scenario". That is, the dynamic process of transition to a chaotic state is depends to the error function. It is determined by the regularity of doubling the number of existing periodicities. In our opinion, this difference is due to the fact that each digit is determined by its distribution of pixel values in the digit array. In confirmation of this, we will consider the maps of dynamic modes when a number is given by an array of 3x5 pixels (Fig. 5).



a) Adam b) AdaMax c) AMSGrad

Fig. 5. Maps of the dynamic learning modes of a three-layer neural network for the printed number "2" with an array of 3x5 pixels, for the optimization methods of learning a) Adam, b) AdaMax, c) AMSGrad, provided that the number of iterations is 10, the optimization parameter beta1=0.9

The decrease in the number of pixels of the digit array caused a change in the spectrum of the existing periodicities. According to Fig. 5, when an array of numbers 3x5 pixels is specified, the process of transition to a chaotic state is carried out gradually by doubling the existing periodicities. That is, a period equal to one is traced – red color, orange – a period equal to 2; yellow – period equal to 3; purple – the period is equal to 7. Comparing the maps of dynamic modes for different arrays of 3x5 and 4x7 numbers, it can be stated that the transition to periodicity with a large period for an array of 3x5 pixels is due to a gradual increase in the period value. That is, the color palette of dynamic mode maps for an array of 3x5 pixels is richer. This may indicate that the representation of a number by an array of 3x5 pixels is more heterogeneous compared to an array of 4x7 pixels. Such heterogeneity of the array affects the learning process of the neural network.

CONCLUSIONS

For a multilayer neural network using Adam, AdamMax, and AMSGrad optimization learning methods, the influence of input array sampling on the learning process was investigated. The conducted studies of maps of the dynamic modes of the error function from the parameters alpha and beta2 (where alpha is the stationary value of the learning step, beta2 is the optimization parameter of the error function, which determines the contribution of the square of the gradient of the error function) of learning, testify to the sensitivity of the learning process to the sampling of the input array.

An increase in the set of the number itself and the variants of their distortions leads to a decrease in the learning error. It was established that the maps of dynamic regimes testify to the existence of a greater number of periodicities for arrays with smaller dimensions. It is shown that the existence of periodicities in the entire interval of changing learning parameters is related to the heterogeneity of the input array. The heterogeneity of the input array contributes to the process of the appearance of local minima, and therefore to more complex behavior of the error function from the learning parameters. That is, the heterogeneity of the input array is a catalyst for the retraining process of the neural network.

REFERENCES

[1] S.O. Subbotin, Neironni merezhi: teoriia ta praktyka: navchalnyi posibnyk [Neural Networks: Theory and Practice. A textbook]. Zhytomyr: O.O. Yevenok Publishing House, 184 p., 2020 (in Ukrainian).

[2] T.O. Savchuk, Ye.O. Yarema, Vykorystannia neironnykh merezh dlia rozpiznavannia symvoliv [The use of neural networks for characters recognition]. Scientific and Technical Journal "Data Recording, Storage & Processing," Institute of Information Recording Problems of the National Academy of Sciences of Ukraine, Kyiv, vol. 7, no. 4, pp. 78-84, 2005. ISSN 1560-9189. (in Ukrainian).

[3] S. Sveleba, I. Katerynchuk, I. Kuno, Y. Shmyhelskyy, N. Sveleba, "The Role of Sample Size in Multilayer Neural Networks," Proceedings of the 16th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, TCSET 2022, 22-26 February 2022, Lviv-Slavske, Ukraine. – pp. 767-770. DOI: 10.1109/TCSET55632.2022.9767025

Vector Ant Algorithm

Fedir Abramov
Department of
General Economic Theory
National Technical University
“Kharkiv Polytechnic Institute”
Kharkiv, Ukraine
Abramov@khpi.edu.ua

Vitaliy Serzhanov
Faculty of Economics
Uzhhorod national university
Uzhhorod, Ukraine
vitaliy.serzhanov@uzhnu.edu.ua

Natalia Reshetniak
Department of
General Economic Theory
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
Natalia.Reshetniak@khpi.edu.ua

Vladimir Zaiats
Department of
Economics and Business
Universitat Pompeu Fabra
Barcelona, Spain
Vladimir.zaiats@upf.edu

Nataliia Volosnikova
Department of
General Economic Theory
National Technical University
“Kharkiv Polytechnic Institute”
Kharkiv, Ukraine
volosnikova@ukr.net

Olga Andreieva
Department of Physics
National Technical University
“Kharkiv Polytechnic Institute”
Kharkiv, Ukraine
Olga.Andreieva@khpi.edu.ua

***Abstract* — This article considers the problem of increasing the efficiency of the use of time by a swarm of robots controlled by an adapted ant algorithm. It is shown that the main reason for the inefficient use of time by the adapted ant algorithm is the use of random walk by the ant algorithm to search for sources of resources and determine the most optimal route for foraging robots from the base to the identified source of the desired resource. To increase the efficiency of time use by a swarm of robots, we proposed a vector ant algorithm, which involves establishing the preferred direction of movement for robots busy searching for sources of resources. In the course of computer modeling, it was shown that the greatest advantages of the proposed vector ant algorithm are manifested when a swarm of robots performs such tasks as searching for the desired resource source and surveying the surrounding area.**

Keywords* — *swarm robotics, ant colony algorithm, nature-inspired algorithm, pheromone memory, optimization.

I. Introduction

The possibility of practical application of any new technology is primarily determined by economic factors, namely the ability to achieve the desired result with the lowest costs. If a new technology achieves a goal at a lower cost than alternative technologies, this new technology will find widespread use. Otherwise, its application will be limited only to those cases when the advantages of this new technology are undeniable.

Swarm robotics has already managed to prove its efficiency and prove the existence of significant advantages, which makes the application of this technology promising for solving certain problems. However, the economic efficiency of using a swarm of robots is determined not so much by the technical aspects of the design of individual robots as members of the swarm, but by the algorithms used to control the operation of the swarm. One such algorithm that was borrowed from collective insects is the adapted ant algorithm.

Similar to other algorithms used to control a swarm of robots (particle swarm optimization [1-3], intelligent water drops [4, 5], gravitational search algorithm [6], and other algorithms [7-9]), the adapted ant algorithm allows a swarm of robots to reveal all its potential advantages while performing a wide range of tasks [10]:

- High level of reliability;
- The possibility of scaling;
- Simplicity of individual members of the swarm, etc.

However, in practice, in addition to the above-mentioned factors, the time factor plays an important role: the swarm of robots must not only be able to cope with the task, but also do it in the shortest possible time, which puts additional requirements on the algorithms for controlling the swarm of robots.

The purpose of this article is to adapt the ant algorithm for the operation of a swarm of robots under conditions of limited time, which is given to complete the task.

II. The Efficiency Coefficient Of The Algorithm

Random walk, a key element of any practical implementation of the ant algorithm, allows to ensure the search of the most optimal route for foraging robots from the base to the identified source of the desired resource. However, this approach turns out to be much less effective in cases where the main task of a swarm of robots is to investigate the presence of certain resources of the entire given territory in the shortest possible time. In the latter case, the low efficiency of random walk is explained, first of all, by the fact that, as a result of random walk, the robots return many times to previously checked areas of the territory. The only consequence of such repeated checks is the loss of time and the slowing down of the entire swarm of robots.

Moreover, due to the fact that every time, when starting a new raid, the robot has to leave the base and pass a small area surrounding the base, the distribution of the number of extra checks by areas is uneven. The largest number of redundant checks falls on those areas that are closer to the base. Areas located on the periphery of the territory are checked relatively few times. If the robot moves through the territory, which is divided into equal square sections, then the dependence of the number of repeated inspections of the sections on their distance from the base (for a swarm of 100 robots after 50,000 steps) will be represented by the following graph in fig. 1.

If we consider as effective only those steps of the robots during which only previously untested areas were explored,

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

then to evaluate the efficiency of the algorithm, we can introduce the efficiency coefficient of the algorithm, which is calculated as the ratio of the number of effective steps to the total number of steps taken by the robots of the swarm, which were controlled by the corresponding algorithm:

$$\eta = \frac{N'}{N}, \qquad (1)$$

where: η is the efficiency coefficient of the algorithm, N' is the number of effective steps, N is the total number of steps.

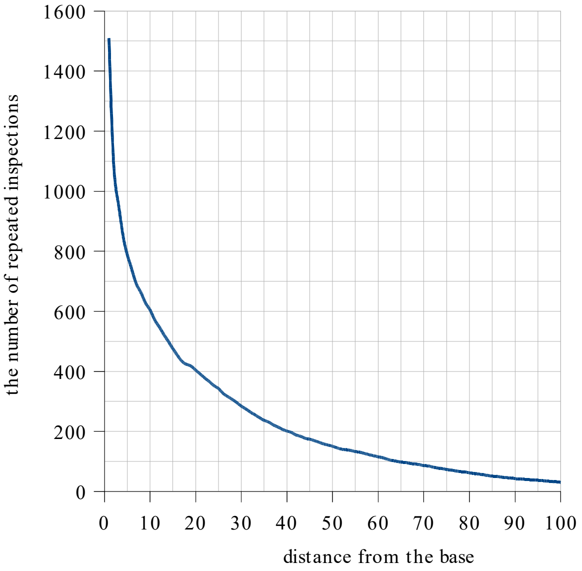


Fig. 1. The dependence of the number of repeated inspections of the sections on their distance from the base.

So, for the adapted ant algorithm (for a swarm of 100 robots after 50,000 steps of the algorithm), the efficiency will be 2.2%. One of the possible ways to improve the efficiency of the adapted ant algorithm is to make changes to this algorithm according to which when the robot selects the next site, to check for the presence of a resource, priority will be given to those neighboring sites that are more distant from the base than the current site in which it is located robot. That is, the procedure for selecting the next site should be changed in the adapted ant algorithm. In the original adapted ant algorithm, it was assumed that the robot during the raid moves through the territory, which is virtually divided into square cells that can be covered by the robot's sensors. Communication between robots and their navigation takes place with the help of virtual pheromone labels, which are applied to the reference and individual maps of pheromone tracks, which are stored, respectively, in the memory of the robots staying at the base and in the memory of each of the robots that went to raid. One of the key performance conditions of the adapted ant algorithm is the simulation of the process of evaporation of pheromone labels applied to the corresponding virtual map, which allows a swarm of robots to "forget" suboptimal routes. To simulate the process of evaporation of pheromone tags, the reference virtual map of pheromone tracks must be updated after each of the robots returns from the raid according to the following rule [11]:

$$I\left(\vec{r}\right) = k \cdot I'\left(\vec{r}\right) + F, \qquad (2)$$

where: $\vec{r}$ – the radius vector of the cell $C\left(\vec{r}\right)$ for which the intensity of the pheromone label is updated; $I\left(\vec{r}\right)$ – the new value of the intensity of the pheromone trail in the cell $C\left(\vec{r}\right)$; $I'\left(\vec{r}\right)$ – previous value of the intensity of the pheromone trace in the cell $C\left(\vec{r}\right)$; k – a coefficient that determines the weight of the previous value of the intensity of the pheromone trail; F – the number of pheromones with which the robot notices each cell it passes.

At each subsequent step, the robot chooses one of the adjacent cells to which it is allowed to move, using the following rule:

$$P\left(\vec{r}+\vec{n_i}\right) = \frac{I\left(\vec{r}+\vec{n_i}\right)+A}{\sum_{j=1}^{4}\left(I\left(\vec{r}+\vec{n_j}\right)+A\right)}, \qquad (3)$$

$$N = \left(\vec{n_1},\ldots,\vec{n_4}\right)$$

$$\vec{n_1} = (0\,;1)$$

$$\vec{n_2} = (1\,;0)$$

$$\vec{n_3} = (0\,;-1)$$

$$\vec{n_4} = (-1\,;0)$$

where: $\vec{r}$ – radius vector of the cell $C\left(\vec{r}\right)$ in which the robot is located; $P\left(\vec{r}+\vec{n_i}\right)$ – the probability of the robot moving to the i-th neighboring cell $C\left(\vec{r}+\vec{n_i}\right)$; $I\left(\vec{r}\right)$ – the intensity of the pheromone trail in the cell $C\left(\vec{r}+\vec{n_i}\right)$; A – a constant that determines the probability of the transition of robots to free search, N – an ordered set of relative coordinates of neighboring cells; $\vec{n_i}$ – the relative coordinates of the i-th neighboring cell $C\left(\vec{r}+\vec{n_i}\right)$.

In a pile with a ban on the robot returning to the cell from which it just moved, this rule for choosing the next cell ensured the preferential movement of the robot along the pheromone track in the direction of the resource source. Thanks to this, the number of steps taken by the robot during the raid is equal to the length of the pheromone track, and the travel time is minimal (provided that the swarm of robots has already found the optimal route). So, at the later stages of the adapted ant algorithm, when the majority of swarm robots are engaged in foraging (the ratio of the number of robots in free raid mode (resource source search mode) to the number of robots engaged in foraging will depend on the coefficient A), this algorithm will be characterized by high efficiency.

The opposite situation will be observed at the initial stages of the adapted ant algorithm, i.e. before the first resource source is discovered. At this stage, there are no pheromone labels on which the robots could orient themselves during their movement, and the rule for choosing the next cell (taking into account the prohibition of returning to the previous cell) takes the following form:

$$P\left(\vec{r}+\vec{n_i}\right)=\frac{1}{3} \quad (4)$$

That is, at the initial stages of the adapted ant algorithm, each of the neighboring cells (except the one from which the transition was made in the previous step) can be selected for the next step with the same probability. And the swarm robots are, in fact, in the mode of random walk, which is the main reason for the low efficiency of the adapted ant algorithm at this stage. Taking into account the above, in order to ensure an accelerated survey of the territory, we proposed a vector ant algorithm, which in the territory survey mode ensures priority movement of robots in the direction from the base to the peripheral areas of the territory being surveyed. The main difference between the vector ant algorithm and the adapted ant algorithm lies in the rule for the robot to select one of the neighboring cells for the next step:

$$P\left(\vec{r}+\vec{n_i}\right)=\frac{I\left(\vec{r}+\vec{n_i}\right)+A+B_i}{\sum_{j=1}^{4}\left(I\left(\vec{r}+\vec{n_j}\right)+A+B_j\right)}, \quad (5)$$

B coefficients are calculated according to the following formulas:

$$B_1=\begin{cases} B\cdot\frac{r_y}{\left|\vec{r}\right|}\cdot\left(1-Min\left(1,\frac{\left|\vec{r}\right|}{L_0}\right)\right), & r_y>0\\ 0, & r_y\le 0\end{cases}$$

$$B_2=\begin{cases} B\cdot\frac{r_x}{\left|\vec{r}\right|}\cdot\left(1-Min\left(1,\frac{\left|\vec{r}\right|}{L_0}\right)\right), & r_x>0\\ 0, & r_x\le 0\end{cases}$$

$$B_3=\begin{cases} -B\cdot\frac{r_y}{\left|\vec{r}\right|}\cdot\left(1-Min\left(1,\frac{\left|\vec{r}\right|}{L_0}\right)\right), & r_y<0\\ 0, & r_y\ge 0\end{cases}$$

$$B_4=\begin{cases} -B\cdot\frac{r_x}{\left|\vec{r}\right|}\cdot\left(1-Min\left(1,\frac{\left|\vec{r}\right|}{L_0}\right)\right), & r_x<0\\ 0, & r_x\ge 0\end{cases}$$

where: B is the coefficient that ensures the priority movement of robots in the direction from the base to the peripheral areas of the territory, L_0 is the coefficient that determines the distance from the base on which the vector ant algorithm operates.

The need to introduce the coefficient L_0 is due to the fact that, as a rule, a swarm of robots is faced with the task of surveying the defined part of the territory, as a result of which, the selection of the priority direction of the robot's movement loses its meaning as the robot approaches the boundary of the defined area. At the beginning of the operation of the vector algorithm (before the first source of the resource is found), there are no trace pheromone labels and the predominant direction of movement of the robots is the direction from the base to the periphery of the territory. After the resource source is found, according to the original adapted ant algorithm, the robot that found it, upon returning to the base, transmits its smoothed (without loops caused by random walk) path from the base to the resource source to the robots at the base and store a reference map of pheromone tracks. This path is applied to the reference map of pheromone paths, which is updated when robot returns to the base according to rule (2).

Further actions of the swarm of robots will depend on the ratio of coefficients *B* and *F*. If the intensity of the trailing pheromone label significantly exceeds the coefficient *B* (that is, *B*<<*F*), then the majority of members of the swarm of robots will move from the base to the detected source of the resource along the corresponding pheromone track. That is, it will switch to the mode of operation of foraging robots that deliver the resource to the base. In the opposite case, when the intensity of the trace pheromone label is significantly lower than the value of the coefficient *B* (i.e., *B*>>*F*), a significant part of the robots will leave the pheromone track and go on a free raid in search of other sources of resources. That is, the swarm will be mainly focused on continued research of this territory. Thus, the mode of operation of a swarm of robots, after the first resource source is found, will be determined by the relative value of the coefficient *B*: the larger this coefficient is, the greater the proportion of robots will be engaged in further exploration of the territories.

If the vector ant algorithm is used to find resources of multiple types, then the number of pheromone types (and corresponding virtual maps) will be equal to the number of resource types, and the simulation of pheromone tag evaporation must occur for each of the available virtual maps [12]. The area of the territory explored by robots can be estimated using the maximum and minimum radii of the explored territory. The first of these indicators - the maximum radius of the studied territory - is defined as the distance from the base of the swarm to the cell that is the most distant from it, which was surveyed by the robots. The second indicator - the minimum radius of the investigated territory - is defined as the distance from the base of the swarm to the least distant cell from it, which has not yet been surveyed by robots. For the case of the original adapted ant algorithm, the maximum and minimum radii of the studied territory, after the completion of 50,000 cycles of the algorithm, are shown in fig. 2. As can be seen from fig. 2, in the case of the original adapted ant algorithm, at the initial stages of the algorithm application, the maximum and minimum radii of the studied territory are close, which is explained by multiple redundant checks of the internal cells of the already studied segment of the territory. As a result, the number of "missed" (that is, unexamined) cells in the corresponding segment is minimal, and only on the periphery of the studied segment do the first "missed" cells begin to appear.

In the case of the vector ant algorithm, due to the reduction of the number of redundant checks, "missed" cells will appear further from the boundary of the studied segment, in the inner part of the studied segment of the territory. As a result, the difference between the maximum and minimum radii of the studied territory will grow, the larger the coefficient *B* is. The feasibility of using the vector ant algorithm will depend on the task set before the swarm of robots, namely on the admissibility of the presence of "missed" cells in the inner part of the studied segment. If the presence of "missed" cells is acceptable, then the efficiency of using the vector algorithm should be determined using the maximum radius of the investigated territory. In the opposite case, using the minimum radius of the investigated territory.

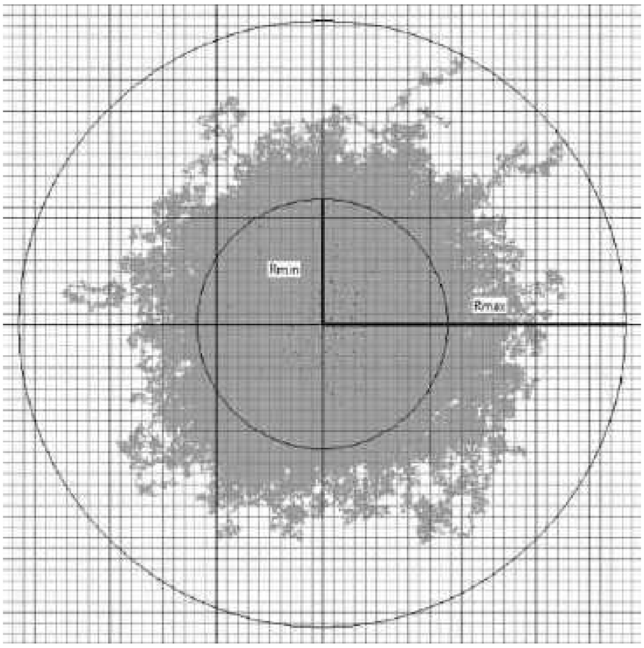


Fig. 2. The maximum and minimum radii of the explored territory.

III. Computer Simulation of the Performance of the Vector Ant Algorithm

In the course of computer simulation, we set up experiments, the purpose of which was:

- comparison of the efficiency of adapted and vector ant algorithms;
- research on the speed of searching for a resource source by a swarm of robots controlled by the vector ant algorithm;
- study of the speed of surveying the surrounding territory by a swarm of robots controlled by the vector ant algorithm.

The simulation was carried out on a map with the size of 1024 by 1024 cells with the base located in the center of the map. The maximum time the robots stayed in the raid was 5,000 cycles. the number of robots in the colony was 100 robots. Coefficients *k*, *F* from equation (2) and coefficients *A*, L_0 from equation (3) in all experiments had the following values:

$$F = 10000\,,$$

$$k = 0.5\,,$$

$$A = 20\,,$$

$$L_0 = 200\,.$$

The dependence of the number of repeated checks of cells on their distance from the base (in the case of the original adapted ant algorithm and the vector ant algorithm) is presented in Fig. 3. In this figure, graphs of the specified dependence for the vector ant algorithm are shown for *B*/*A* equal to 0.25; 0.50; and 0.75 and for the adapted ant algorithm, which can be considered as a case of the vector ant algorithm with a coefficient *B* equal to zero (*B*/*A*=0). As can be seen from the graphs, with the growth of the coefficient *B*, the number of repeated checks of the cells located close to the base decreases, which indicates less time loss by the robot swarm controlled by the vector ant algorithm.

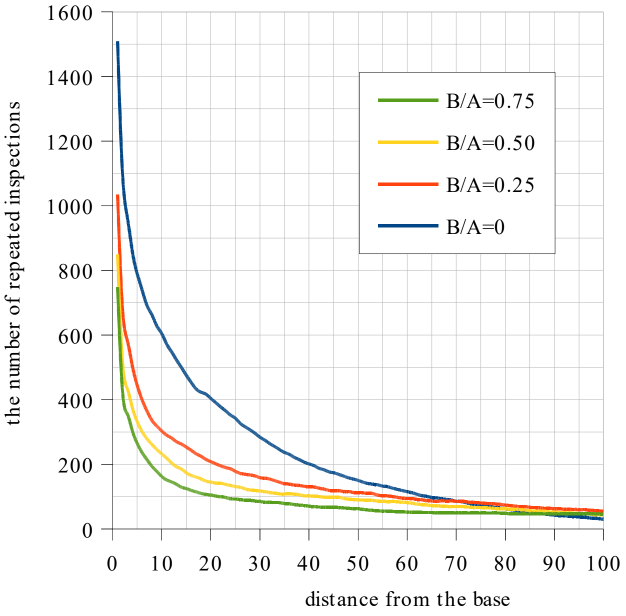


Fig. 3. The dependence of the number of repeated checks of cells on their distance from the base.

The location of the swarm member robots and the area of the surveyed area for the cases of the adopted ant algorithm and the vector ant algorithm, obtained as a result of computer simulation, for *B*/*A* values equal to 0; 0.25; 0.50; and 0.75 after 100,000 cycles after the start of these algorithms, shown in Fig. 4, 5, 6 and 7.

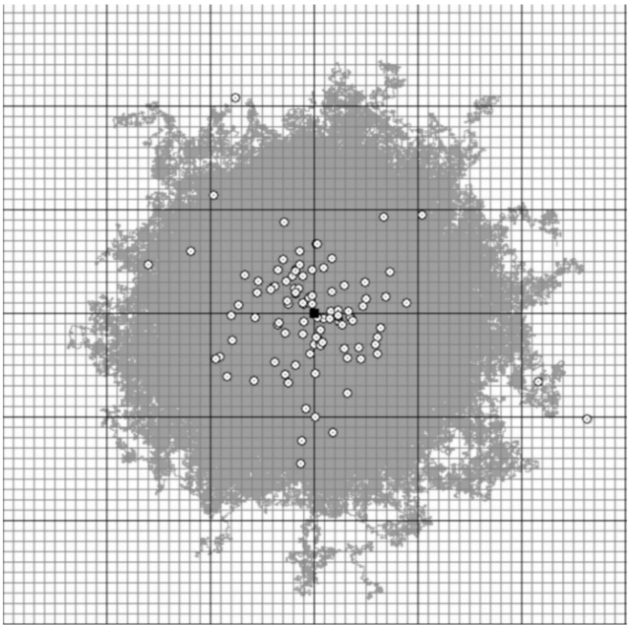


Fig. 4. The location of the swarm member robots and the area of the surveyed area for the cases of the adopted ant algorithm after 100,000 cycles after the start of the algorithm.

Graphs of efficiency vs. time (in cycles) for the adapted ant algorithm and for the vector ant algorithm for *B*/*A* values equal to 0; 0.25; 0.50; and 0.75 are shown in Fig. 8 (this and the following graphs do not show values for time intervals shorter than the maximum time the robots stay in the raid, since the pheromone map remains not updated).

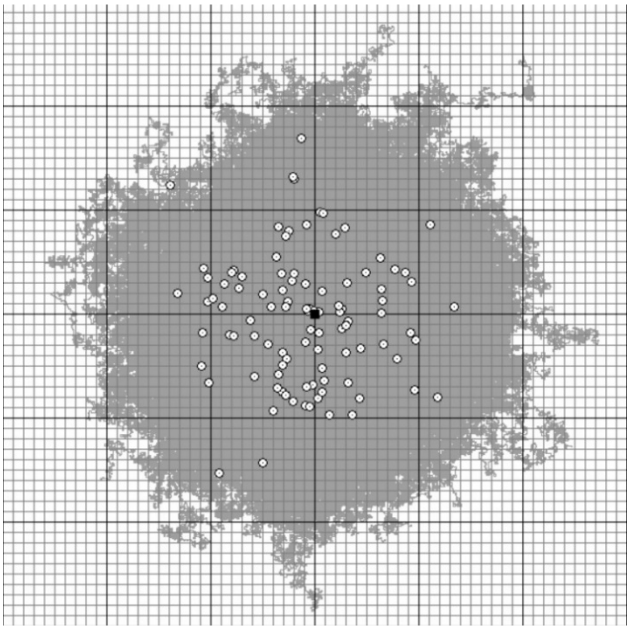


Fig. 5. The location of the swarm member robots and the area of the surveyed area for the cases of the vector ant algorithm, for *B*/*A* = 0.25 after 100,000 cycles after the start of the algorithm.

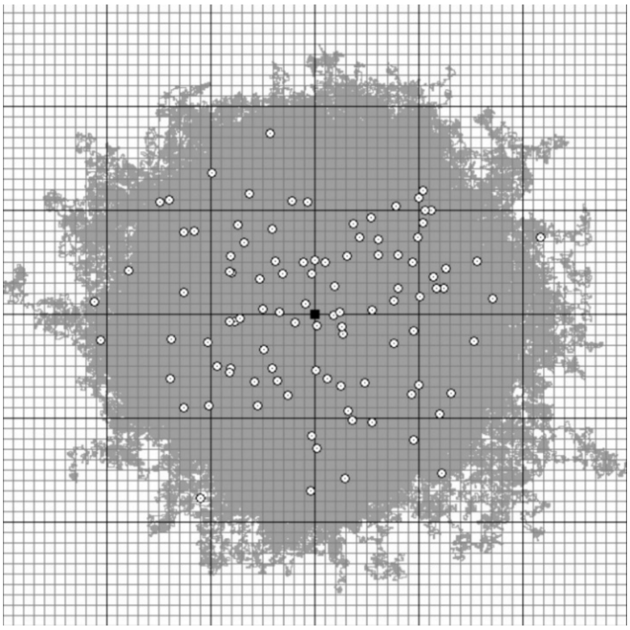


Fig. 6. The location of the swarm member robots and the area of the surveyed area for the cases of the vector ant algorithm, for *B*/*A* = 0.50 after 100,000 cycles after the start of the algorithm.

For both algorithms, the efficiency decreases over time due to the increase in the number of retests, but for any time interval that has passed since the algorithms started, the utility of the vector ant algorithm is higher than the utility of the adaptive ant algorithm. In these figures, unexamined cells are marked in white, and examined cells are marked in gray. The base of the swarm, which is located in the center of the figure, is marked with a black square, and the robots are marked with white circles.

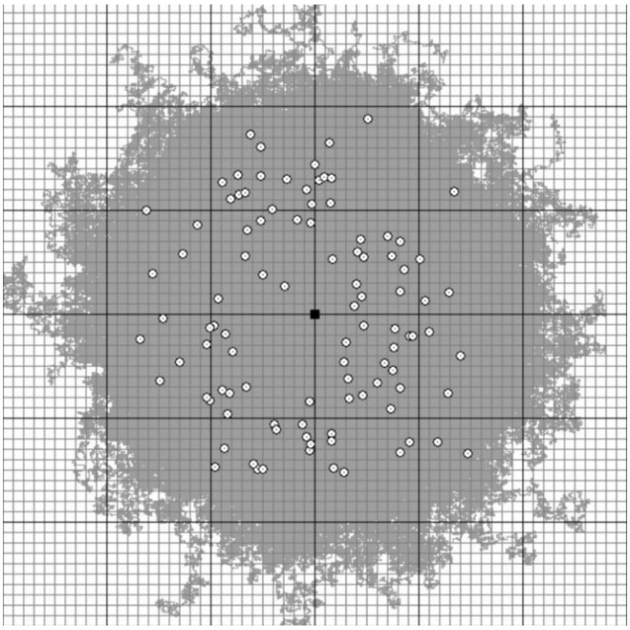


Fig. 7. The location of the swarm member robots and the area of the surveyed area for the cases of the vector ant algorithm, for *B*/*A* = 0.75 after 100,000 cycles after the start of the algorithm.

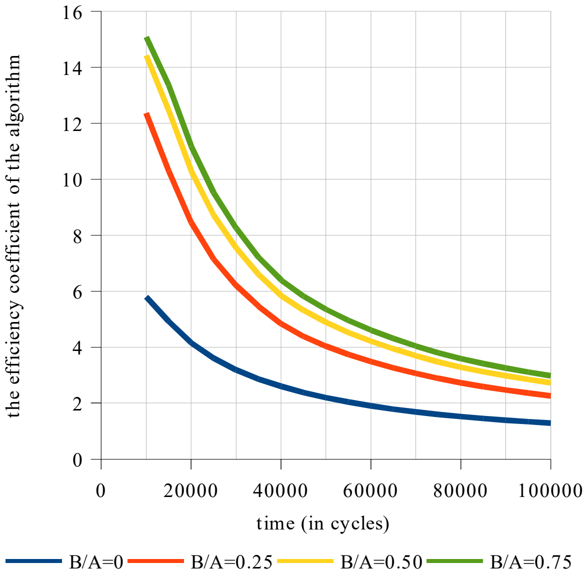


Fig. 8. Efficiency vs. time (in cycles) for the adapted ant algorithm and for the vector ant algorithm.

Graphs of the dependence of the search time for a resource source located 150 cells from the base on the *B*/*A* ratio are shown in Fig. 9. As can be seen from this graph, for any value of *B*/*A*, the speed of finding a resource source, in the case of the vector ant algorithm, is greater than in the case of the adaptive ant algorithm. Graphs of dependence of the maximum radius of the examined territory R_{max} from time, for the vector ant algorithm for *B*/*A* values equal to 0.25; 0.50; and 0.75 and for the adapted ant algorithm (*B*/*A*=0) are shown in Fig. 10. A similar dependence for the minimum radius of the surveyed area R_{min} (for the same values of *B*/*A*) is shown in Fig. 11. As you can see from the last graph, as the value of *B*/*A* increases, the minimum radius of the surveyed area R_{min} also increases. However, for small time intervals since the start of the algorithm, the minimum surveyed area radius R_{min} for the vector ant algorithm is smaller than R_{min} for both the adapted and the vector ant algorithm with a smaller *B*/*A* value.

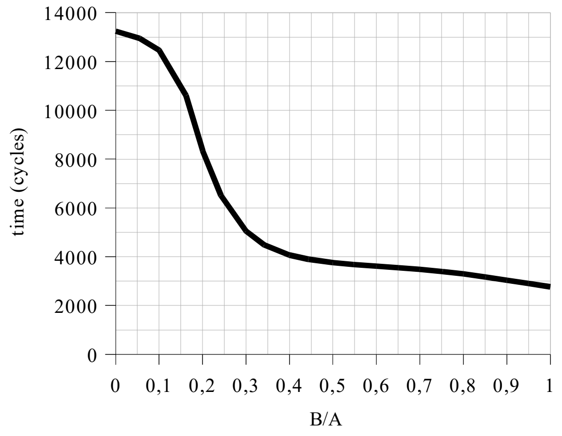


Fig. 9. The dependence of the search time for a resource source (in clocks) located 150 cells from the base.

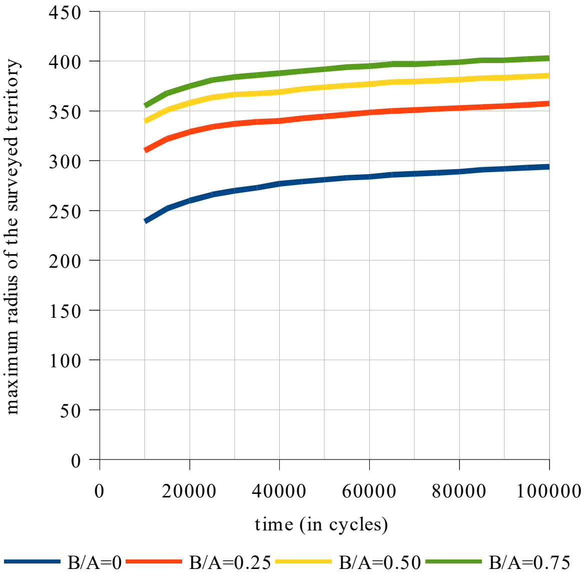


Fig. 10. Dependence of the maximum radius of the examined territory R_{max} from time (in clocks).

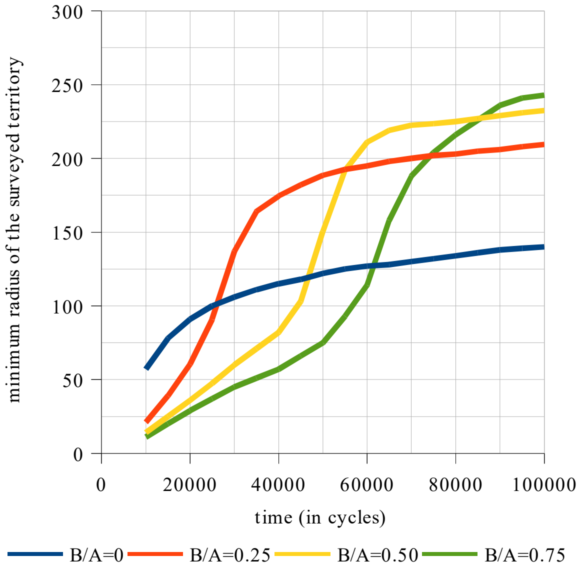


Fig. 11. Dependence of the minimum radius of the examined territory R_{max} from time (in clocks).

As a result, the choice of the value of B/A will depend both on the area of the territory to be explored by the swarm of robots, and on the time given to the swarm of robots to perform the task of surveying the territory.

Conclusions

The computer simulation proved that the proposed vector ant algorithm is workable and effective. Compared to the adapted ant algorithm, the vector ant algorithm is characterized by a higher value of the coefficient of useful action for any time interval that has passed since the start of the algorithms. Also, due to the reduction of the number of redundant checks of the areas of the studied territory, compared to the adapted ant algorithm, the vector ant algorithm takes less time to search for the source of the resource, and also allows you to explore a larger area than the adapted ant algorithm in a given time.

References

[1] Hamami, M.G.M., Ismail, Z.H. A Systematic Review on Particle Swarm Optimization Towards Target Search in The Swarm Robotics Domain. Archives of Computational Methods in Engineering (2022). https://doi.org/10.1007/s11831-022-09819-3

[2] Nayak, J., Swapnarekha, H., Naik, B. et al. 25 Years of Particle Swarm Optimization: Flourishing Voyage of Two Decades. Archives of Computational Methods in Engineering 30, 1663–1725 (2023). https://doi.org/10.1007/s11831-022-09849-x

[3] Atyabi, A., Phon-Amnuaisuk, S. & Ho, C.K. Applying Area Extension PSO in Robotic Swarm. Journal of Intelligent and Robotic Systems 58, 253–285 (2010). https://doi.org/10.1007/s10846-009-9374-2

[4] Rao, D.C., Kabat, M.R., Das, P.K. et al. Hybrid IWD-DE: A Novel Approach to Model Cooperative Navigation Planning for Multi-robot in Unknown Dynamic Environment. Journal of Bionic Engineering 16, 235–252 (2019). https://doi.org/10.1007/s42235-019-0020-9

[5] Salmanpour, S., Monfared, H. & Omranpour, H. Solving robot path planning problem by using a new elitist multi-objective IWD algorithm based on coefficient of variation. Soft Computing 21, 3063–3079 (2017). https://doi.org/10.1007/s00500-015-1991-z

[6] Das, P.K., Behera, H.S., Jena, P.K. et al. An intelligent multi-robot path planning in a dynamic environment using improved gravitational search algorithm. International Journal of Automation and Computing 18, 1032–1044 (2021). https://doi.org/10.1007/s11633-016-1019-x

[7] Li, G., Chou, W. Path planning for mobile robot using self-adaptive learning particle swarm optimization. Science China Information Sciences 61, 052204 (2018). https://doi.org/10.1007/s11432-016-9115-2

[8] Ayari, A., Bouamama, S. A new multiple robot path planning algorithm: dynamic distributed particle swarm optimization. Robotics and Biomimetics 4, 8 (2017). https://doi.org/10.1186/s40638-017-0062-6

[9] Soleimanpour-moghadam, M., Nezamabadi-pour, H. A multi-robot task allocation algorithm based on universal gravity rules. International Journal of Intelligent Robotics and Applications 5, 49–64 (2021). https://doi.org/10.1007/s41315-020-00158-9

[10] F. V. Abramov, A. Andreiev, O. Andreieva, “Implementation of an algorithm for searching for missing units of a swarm of robots, controlled by an adapted ant algorithm,” 2021 IEEE 2nd KhPI Week on Advanced Technology (KhPIWeek), 2021 - Conference Proceedings, pp. 336-340. doi.org/10.1109/KhPIWeek53812.2021.9570007

[11] F. Abramov, O. Andreieva and O. Andreiev, "Adaptation of the Ant Algorithm to Control a Robot Swarm," 2020 IEEE KhPI Week on Advanced Technology (KhPIWeek), 2020, pp. 500-503. doi.org/10.1109/KhPIWeek51551.2020.9250088

[12] F. V. Abramov, “Implementation of an Adapted Ant Algorithm in the Presence of Substitute and Complementary Resources. Modeling the Behavior of the Manufacturer,” 2021 IEEE 2nd KhPI Week on Advanced Technology (KhPIWeek), 2021 - Conference Proceedings, pp. 341-345. doi.org/10.1109/KhPIWeek53812.2021.9570095

Increasing the Efficiency of Robot Swarm Navigation with the Help of Virtual Pheromone Direction Labels

Fedir Abramov
Department of
General Economic Theory
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
Abramov@khpi.edu.ua

Vitaliy Serzhanov
Faculty of Economics
Uzhhorod national university
Uzhhorod, Ukraine
vitaliy.serzhanov@uzhnu.edu.ua

Oleksandr Andreiev
Department of Physics
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
Oleksandr.Andreiev@khpi.edu.ua

Olga Andreieva
Department of Physics
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
Olga.Andreieva@khpi.edu.ua

Tetiana Diachenko
Department of
General Economic Theory
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
tatyana.oet@gmail.com

Iana Maksymenko
Department of
General Economic Theory
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
maksimenko.yana@gmail.com

***Abstract* — This article deals with the problem of improving the navigation efficiency of a swarm of robots controlled by an adapted ant algorithm. It is shown that in the presence of an extensive network of virtual pheromone tracks, the efficiency of a swarm of robots can be reduced due to the formation of closed loops by pheromone tracks, through which robots can wander during their movement to the desired resource source. It is also shown that the problem of increasing the efficiency of navigation of a swarm of robots can be solved with the help of virtual pheromone direction labels. In the course of computer simulation, the performance of the proposed oriented ant algorithm and its ability to increase the efficiency of navigation of a swarm of robots were confirmed.**

Keywords* — *swarm robotics, nature-inspired algorithm, navigation, efficiency, pheromone memory.

I. Introduction

The rapid progress of swarm robotics, which has been observed in recent times, is primarily due to the improvement of algorithms for controlling the work of a swarm of robots. Today, various algorithms are used to control a swarm of robots. A special place among these algorithms is occupied by the ant algorithm, which allows a swarm of robots to efficiently perform such tasks as finding an effective route to the identified resource source and surveying the given territory.

Compared to all other algorithms for controlling a swarm of robots [1-8], the main difference between all existing varieties of the ant algorithm is the use of individual units of the swarm system of pheromone trails or their analogues. The system of pheromone trails plays two main functions: it ensures the dissemination among swarm members of information about detected sources of resources, and it also ensures navigation of swarm member robots. And it is this system of pheromone trails that provides the ant algorithm with all its advantages. However, the same system of pheromone trails can become a source of many problems that can negatively affect the efficiency of the swarm of robots controlled by the ant algorithm. The biggest of such problems is that in the initial stages of finding the optimal route to the identified resource source, there may be several alternative pheromone trails that may have many crossing points. As a result, the swarm of robots is faced not with one pheromone path, but with a labyrinth of pheromone paths, which significantly complicates navigation.

The purpose of this article is to improve the efficiency of navigation and operation of a swarm of robots, the members of which are guided by an adapted ant algorithm in the presence of several alternative pheromone paths to the detected resource source.

II. The Oriented Ant Algorithm

Any implementation of the ant algorithm, including the adapted ant algorithm [9], involves the use by robots that have successfully reached the resource source of pheromone tags to distribute information about the location of the source among the members of the swarm. Due to the fact that the resource source can be independently found by several robots, several alternative pheromone paths to the resource source can exist at the same time. In its further work, the swarm of robots chooses the most efficient path to the resource source among the available alternatives, focusing on the intensity of pheromone tags. However, the extensive network of pheromone trails formed at the initial stages of the ant algorithm creates favorable conditions for the formation of loops and intersections of pheromone trails, as a result of which some robots may return to the base without reaching the corresponding resource source, or start wandering in a circle until time being in a raid runs out. Accordingly, the formation of loops on the pheromone trails leads to a decrease in the efficiency of the adapted ant algorithm, because some of the robots, due to wandering through the loops, spend more time reaching the resource source than is predicted by any of the available pheromone trails, and some of them return to the base completely empty.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

For further analysis, let's first consider the process of forming a branched network of pheromone trails. At the same time, we will present the network of pheromone trails along which swarm robots can move in the form of an undirected graph, the edges of which are represented by unbranched segments of pheromone trails, and the nodes are branching points of pheromone trails. Thus, for a robot moving from the base (vertex *A*) to the resource source (vertex *B*) along the edge *AB*, there is a non-zero probability at any point of the pheromone path (vertex *C*) to go into free raid mode (Fig. 1). For the robot that has entered the free raid mode at vertex *C*, there are three possible options for ending the raid, which will have different effects on the efficiency of the adapted ant algorithm.

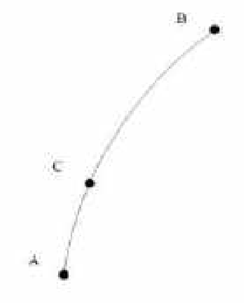
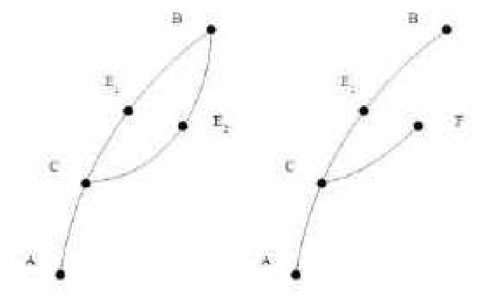


Fig. 1. The process of forming a branched network of pheromone trails.

First, a robot that has entered a free raid can independently find a way to a given (Fig. 2a) or other resource source (Fig. 2b) different from what was already known to the swarm of robots and was noticed on the reference map of pheromone trails. Upon returning to base, this new path will be added to the reference map of pheromone trails maintained by the base robots. Subsequently, a choice between alternative paths (between *AC*, *CE*$_1$, *E*$_1$*B* and *AC*, *CE*$_2$, *E*$_2$*B*, in the case of the same source, or between *AC*, *CE*$_1$, *E*$_1$*B* and *AC*, *CF*, in the case of two different resource sources) will be made by swarm according to the alternative ant algorithm.

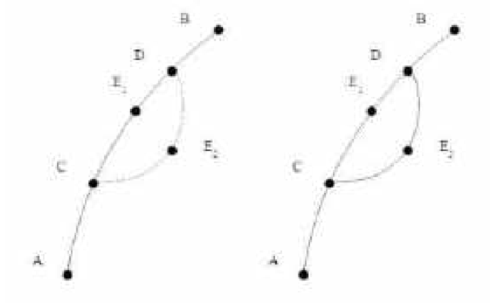


a) b)

Fig. 2. Alternative paths

Second, a robot that has entered a free raid may not find an alternative path and, after reaching the maximum time spent in the raid, it will return to the base without making any changes to the reference map of pheromone trails. Despite the fact that such a raid is ineffective, it does not reduce the effectiveness of the adapted ant algorithm.

Thirdly, after transitioning to a free raid, at vertex *C*, the robot can return to the pheromone path at another point (vertex *D*) (Fig. 3a). The consequences of this course of events for the efficiency of the adapted ant algorithm will depend on which further direction of movement will be chosen by the robot: in the direction of the base (in the direction of vertex *C*) or in the direction of the source of the resource (in the direction of vertex *B*). In the event that the robot moves in the direction of vertex *C*, the robot will return to the base without finding a path to the source and thus without making any changes to the reference map of pheromone trails. That is, the consequences will be similar to the second variant of the end of the raid. If the robot moves to point *B*, the robot will reach the resource source and, upon returning to the base, will add a new alternative path to the reference map of pheromone trails, consisting of edges: *AC*, *CE*$_2$, *E*$_2$*D*, *DB* (Fig. 3b). In the latter case, a closed loop *CE*$_2$*DE*$_1$*C* is formed.



a) b)

Fig. 3. Forming a loop.

Of all the cases discussed above, only the last one creates conditions that can significantly complicate the navigation of robots along the available pheromone trails. Thus, during the first passage of the vertex *C* closest to the base, the choice of the direction of further movement by the robot will not affect the effectiveness of the further work of the swarm of robots. After all, no matter which edge is chosen by the robot (*CE*$_1$ or *CE*$_2$), after passing vertex *C*, it will still continue to move in the direction of the corresponding resource source — vertex *B*. A different situation will be observed after the robot reaches vertex *D*. In this vertex, the robot can choose an edge *DB* and continue its movement in the desired direction to the source of the resource, or it can choose the edge *E*$_1$*D* (*E*$_2$*D*) and start moving in the opposite direction — from the source of the resource to the base. Having chosen the edge *E*$_1$*D* (*E*$_2$*D*), the robot will move to the vertex *C*, where it will complete the circle *CE*$_2$*DE*$_1$*C* and will have to choose the further direction of its movement again. Having chosen the edge *AC*, the robot will return to the base ahead of time without reaching the source of the resource. At the same time, depending on the implementation of the algorithm, upon reaching the base, the robot will either start a new raid, or continue this raid by switching to free raid mode. Having selected the edge *CE*$_2$ (*CE*$_1$), the robot will start a new circle of movement along the cycle *CE*$_2$*DE*$_1$*C*. Such movement in a circle will continue until:

- the robot will select the edge *DB* at the vertex *D* for further movement;
- the robot will select edge *AC* at vertex *C* for further movement;
- the maximum time the robot stays in the raid will not expire.

In any case, the maximum time the robot wanders along the available pheromone trails will not exceed the maximum time the robot stays in a free raid. At the same time, the presence of an extensive network of pheromone trails does not have any effect on the duration of the robot's return journey from the resource source to the base. After all, according to the adapted ant algorithm, after reaching the resource source (or after exhausting the maximum time spent in the raid), the robot smooths the path it has traveled, discarding all the loops formed as a result of random wandering and returns to the base along this smoothed path. Therefore, no matter how many complete circles the robot makes along the closed loops formed by the pheromone trails, the robot will return to the base on a smooth path. From the point of view of the possibility of organizing the effective delivery of the resource to the base, that is, maximizing the number of effective raids carried out per unit of time, the most important consequences of the robot wandering through the existing network of pheromone trails are:

- an increase in the average duration of a successful raid, which occurs as a result of the robot passing through extra loops while moving to the resource source;
- an increase in the share of unproductive raids that occur due to the premature return of robots to the base, or the robot's ineffective wandering along the pheromone paths before the maximum time of the robot's stay in the raid is exhausted.

Both of these factors are capable of reducing the number of effective raids carried out during a certain period of time. Thus, an increase in the average duration of a successful raid means that in the same period of time the robot will have time to make a smaller number of successful raids, and therefore will be able to deliver a smaller number of units of the required resource to the base. An increase in the share of unsuccessful raids will mean that fewer robots will be involved in the process of delivering resources. Thus, as can be seen from the above, both the formation of a branched network of pheromone trails and the wandering of robots through the loops formed by these trails are a direct consequence of the peculiarities of the implementation of the adapted ant algorithm. And if the first component of the problem – the formation of an extensive network of pheromone trails – cannot be eliminated, because this will lead to the inability of the algorithm to search for the most efficient path to the source of the resource, then the second component of the problem can be solved by making certain changes in the implementation of the adapted ant algorithm.

To solve the considered problem, we proposed an oriented ant algorithm, which is an improved version of the adapted ant algorithm. In this algorithm, preventing robots from wandering through loops of pheromone trails is achieved by the fact that not only the intensity of use of this route by the rest of the robots, but also the direction of movement along this trail to the source of the resource is encoded in the pheromone tags. For this purpose, each cell of the pheromone trail map contains information not only about the intensity of use of the route passing through it, but also information about the direction of movement. This is achieved by using two types of pheromone tags: lane usage intensity pheromone tags and direction pheromone tags. Pheromone usage intensity labels, as in the adapted ant algorithm, are designed to identify the most optimal routes from the base to the resource source. Their intensity reflects how often this route is used by a swarm of robots. The purpose of pheromone direction labels is to prevent robots from wandering in a circle in a labyrinth of pheromone trails. On the other hand, pheromone direction labels should not prevent the search for alternative routes from the base to the given source or other sources of the resource. The specified requirements can be fulfilled if the pheromone direction tags will contain information about the neighboring cells from which the transition to this cell can be made.

Thanks to pheromone direction labels, the network of pheromone trails along which swarm robots can move can be represented as a directed graph. Accordingly, in order to prevent the robot from moving in the opposite direction, when the robot selects one of the neighboring cells for the next step, from the list of neighboring cells to which the transition is allowed, those cells whose information is contained in the pheromone label of the direction of the cell in which robot is located must be excluded. Thus, the oriented ant algorithm is reduced to the following. Virtual pheromone labels of trail usage intensity and pheromone direction labels are used for navigation and dissemination of information about detected resource sources. During the raid, the robot uses its individual map of pheromone tags. The reference map of pheromone labels is kept by the robots at the base. The latter is updated every time any robot returns to the base. The territory on which the robot moves is divided into virtual identical square sections. Leaving the base, the robot moves from the current cell to one of the neighboring cells. The robot makes one step per unit of time. The rule according to which the robot chooses for the next step one of the neighboring cells to which the transition is allowed can be presented in the following form:

$$P\left(\vec{r}+\vec{n_i}\right)=\begin{cases}\dfrac{I\left(\vec{r}+\vec{n_i}\right)+A}{\sum\limits_{\substack{j=1\\ \{\vec{n_i}\}\not\subset D(\vec{r})}}^{m}\left(I\left(\vec{r}+\vec{n_j}\right)+A\right)}, & \{\vec{n_i}\}\not\subset D(\vec{r})\\ 0, & \{\vec{n_i}\}\subset D(\vec{r})\end{cases} \tag{1}$$

$$N=\left(\vec{n_1},\ldots,\vec{n_4}\right)$$

$$\vec{n_1}=(0\,;1)$$

$$\vec{n_2}=(1\,;0)$$

$$\vec{n_3}=(0\,;-1)$$

$$\vec{n_4}=(-1\,;0)$$

where: $\vec{r}$ – radius vector of the cell $C\left(\vec{r}\right)$ in which the robot is located; $P\left(\vec{r}+\vec{n_i}\right)$ – the probability of the robot moving to the *i*-th neighboring cell $C\left(\vec{r}+\vec{n_i}\right)$; $I\left(\vec{r}\right)$ – pheromone labels of trail usage intensity in the cell $C\left(\vec{r}+\vec{n_i}\right)$; $D\left(\vec{r}\right)$ – the pheromone label of the direction in the cell $C\left(\vec{r}+\vec{n_i}\right)$; A – a constant that determines the probability of the transition from work to free search, N – an ordered set of relative coordinates of neighboring cells; $\vec{n_i}$ – the relative coordinates of the *i*-th neighboring cell $C\left(\vec{r}+\vec{n_i}\right)$.

The robot continues to move from one cell to another until a cell with a resource source located there is found, or until the maximum time the robot can stay in the raid is exhausted. If the raid was successful (the source of the resource was found), then upon returning to the base, the robot transmits its smoothed path to the robots supporting the reference map of pheromone trails for its update. The latter, having received information about the new path, check it for compliance with the pheromone direction labels of the reference map. The path of the robot is added to the reference map of pheromone trails, if the proposed path does not involve transitions between neighboring cells in a direction that is forbidden according to the available pheromone direction labels, that is, it meets the following condition:

$$\left(\vec{tr}_{i-1} - \vec{tr}_i\right) \notin D\left(\vec{tr}_i\right), \text{ for all } 0 < i < N_{tr} \tag{2}$$

where: $\vec{tr}_i$ – coordinates of the cell visited by the robot at the i-th step of the raid; N_{tr} is the maximum duration of the raid.

If the path can be added to the reference map, then the value of the pheromone labels of the intensity of use and the pheromone labels of the direction is pre-updated according to the following formulas:

$$I\left(\vec{r}\right) = \begin{cases} kI'\left(\vec{r}\right) + F, & I'\left(\vec{r}\right) \geq I_{min} \\ 0, & I'\left(\vec{r}\right) < I_{min} \end{cases} \tag{3}$$

$$D\left(\vec{r}\right) = \begin{cases} D'\left(\vec{r}\right), & I\left(\vec{r}\right) > 0 \\ 0, & I\left(\vec{r}\right) = 0 \end{cases} \tag{4}$$

$$D\left(\vec{r} + \vec{n}_i\right) = \begin{cases} D'\left(\vec{r} + \vec{n}_i\right), & D\left(\vec{r}\right) > 0 \\ D'\left(\vec{r} + \vec{n}_i\right) \setminus \left\{-\vec{n}_i\right\}, & D\left(\vec{r}\right) = 0 \end{cases} \tag{5}$$

where: $I\left(\vec{r}\right)$ – the new value of pheromone labels of trail usage intensity in the cell; $I'\left(\vec{r}\right)$ – previous value of the intensity of the pheromone trail; k – a coefficient that determines the weight of the previous value of the intensity of the pheromone trail; F – the number of pheromones with which the robot notices each cell it passes; $D\left(\vec{r}\right)$ – the new value of the pheromone direction tag in the cell; $D'\left(\vec{r}\right)$ – the previous value of the pheromone direction tag in the cell.

After that, pheromone direction labels are added to the reference map:

$$D\left(\vec{tr}_i\right) = D'\left(\vec{tr}_i\right) \cup \left\{\vec{tr}_{i-1} - \vec{tr}_i\right\} \tag{6}$$

If the proposed path does not meet the condition (2), then it is rejected, and the reference map remains unchanged. Checking the correctness of pheromone direction labels along a new path, before adding it to the reference map by the robot, is a necessary condition for the performance of the proposed algorithm. The need for this check is due to the following.

A robot moving along a pheromone trail or in a free raid makes a transition to the next cell according to rule (1), which prohibits transition to those cells whose pheromone direction label indicates the opposite direction of movement between these cells and the cell in which the robot is located. However, when checking the admissibility of a transition, the robot can only use the information contained in its own instance of the pheromone trail map stored in its RAM. During the time when the robot is in the raid, other robots may have time to return to the base and update the reference map of pheromone trail stored at the base. As a result, the transition between individual cells along the path along which the robot moved during the free raid, at the time of returning to the base, may be prohibited. Before each subsequent raid, the robot replaces its own copy of the pheromone map with the reference map.

III. Computer Simulation of the Performance of the Oriented Ant Algorithm

To simulate the effect of the presence of alternative routes from the base to the resource source, we used artificial pheromone trails from the base A to the resource source B. The main pheromone trail consisted of edges AC, CE_1, E_1D and DB in Fig. 4. (shown in the figure by a solid green bold line).

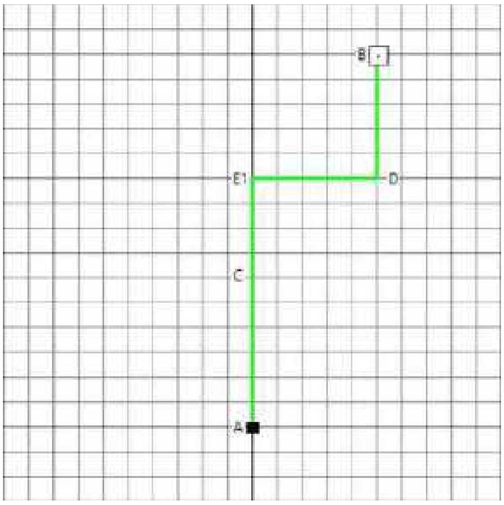


Fig. 4. The main pheromone trail.

The additional pheromone trail consisted of edges AC, CE_2, E_2D, and DB of Fig. 5 (shown in the figure by a dashed black bold line).

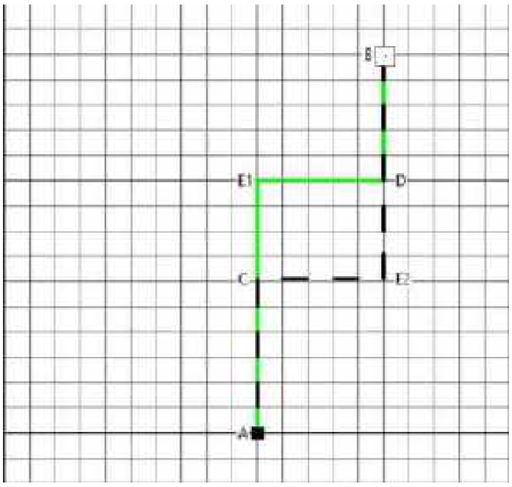


Fig. 5. The additional pheromone trail.

Edges CE_1 and E_1D, which were only part of the main trail, were marked with pheromone labels with an intensity of 10,000 units. The intensity of these pheromone labels remained unchanged throughout the experiment. Edges CE_2 and E_2D, which were only part of the alternative trail, were marked with pheromone labels with different intensities from 0 to 10,000 units. The intensity of the pheromone labels of the edge's AC and DB, which were part of both routes at the same time, was equal to the sum of the intensities of the edges CE_1 and CE_2. The length of both pheromone trails (AC, CE_1, E_1D, DB and AC, CE_2, E_2D, DB) was the same and consisted of 200 cells. If, due to wandering along pheromone paths, the robot returned to the base prematurely (without first visiting the resource source), the current raid was considered completed and ineffective, and the robot went on a new raid. Since the ability of the robots to deviate from it while moving along the pheromone path and go into the free raid mode embedded in the adapted ant algorithm could affect the simulation results, to prevent this, the coefficient A was set equal to 0. Due to this, the robots could not deviate while moving from the pheromone trail, and the increase in the average time spent by the robot in a productive raid and the reduction in the number of productive raids were caused solely by wandering robots along the closed cycles formed by the pheromone paths. The graph of the dependence of the average duration of a successful raid on the ratio of the intensities of the main and alternative routes I_{ACE2DB}/I_{ACE1DB}, after 50,000 cycles of operation of the algorithm, is shown in Fig. 6.

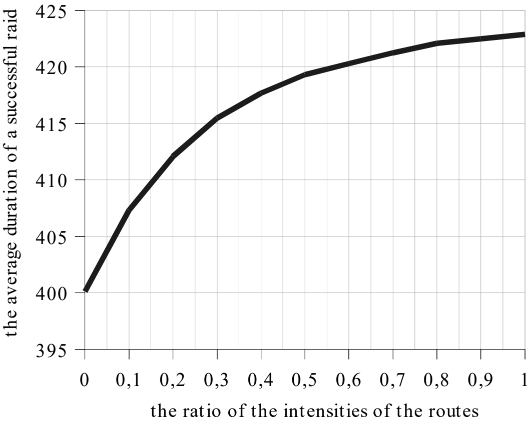


Fig. 6. The dependence of the average duration of a successful raid on the ratio of the intensities of the main and alternative routes I_{ACE2DB}/I_{ACE1DB}, after 50,000 cycles.

As can be seen from the given graph, the average duration of a successful raid is minimal in the absence of an alternative route (I_{ACE2DB}/I_{ACE1DB}=0) and is equal to twice the length of the pheromone trail. As the intensities of the pheromone labels of the alternative route increase, the average duration of a successful raid increases nonlinearly, due to the fact that a certain part of the robots reaches the resource source only after several extra rounds of the CE_2DE_1C cycle. However, the effectiveness of a swarm of robots should not be determined by the average duration of a successful raid, but by the number of raids carried out within a certain period of time. After all, the reason for the reduction in the number of successful raids carried out during a certain period of time depends not only on the average duration of a successful raid, but also on the increase in the share of unsuccessful raids. Graphs of the dependence of the number of successful raids on the duration of the original adapted ant algorithm for cases where the intensity ratio of the main and alternative I_{ACE2DB}/I_{ACE1DB} routes is equal to 0; 0.25; 0.5 and 1 are shown in fig. 7. As can be seen from the graphs, the number of successful raids increases with the increase in the running time of the original adapted ant algorithm for all values of I_{ACE2DB}/I_{ACE1DB}. However, with the growth of I_{ACE2DB}/I_{ACE1DB}, the rate of growth of successful raids decreases. The location of the robots of the swarm members after 50,000 cycles of the original adapted ant algorithm for the ratio of the intensities of the main and alternative routes I_{ACE2DB}/I_{ACE1DB} equal to 0.25 and 0.75 is shown, respectively, in Fig. 8 and 9. In these figures, the base of the swarm is marked with a black square, and the source of the resource to which the artificial pheromone trails are laid is marked with a white square. Robots are marked with white circles.

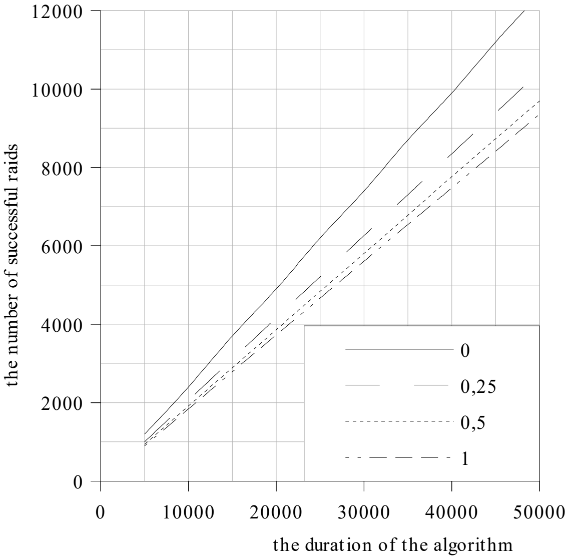


Fig. 7. The dependence of the number of successful raids on the duration of the original adapted ant algorithm.

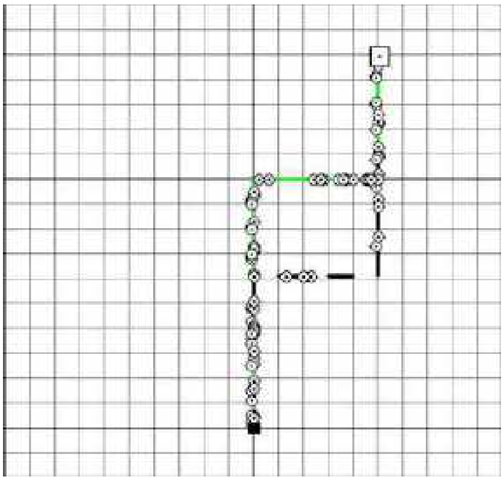


Fig. 8. The location of the robots of the swarm members after 50,000 cycles of the original adapted ant algorithm for the I_{ACE2DB}/I_{ACE1DB} equal to 0.25.

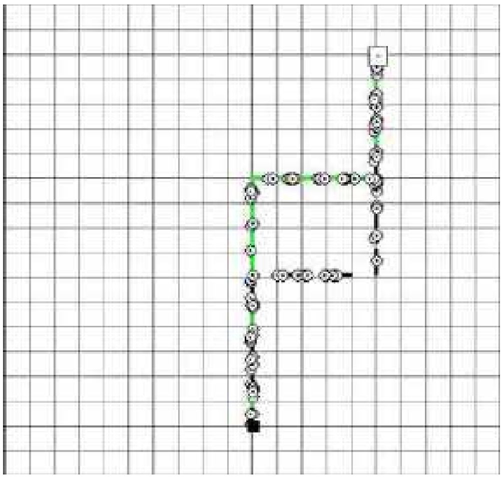


Fig. 9. The location of the robots of the swarm members after 50,000 cycles of the original adapted ant algorithm for the I_{ACE2DB}/I_{ACE1DB} equal to 0.75.

Graphs of the dependence of the average duration of a successful raid on the ratio of the intensities of the main and alternative routes I_{ACE2DB}/I_{ACE1DB}, after 50,000 cycles of the original adapted ant algorithm (shown in the figure by a solid bold line) and the oriented ant algorithm (shown in the figure by a dashed bold line), are shown in fig. 10. It can be seen from the graphs that, in the case of using the oriented ant algorithm, the average duration of a successful raid remains unchanged for all values of I_{ACE2DB}/I_{ACE1DB}. At the same time, in this case, the duration of a successful raid is equal to the average duration of a successful raid for the case of having only one route to the resource source (I_{ACE2DB}/I_{ACE1DB} =0), i.e. it is determined solely by the length of the pheromone path to the resource source.

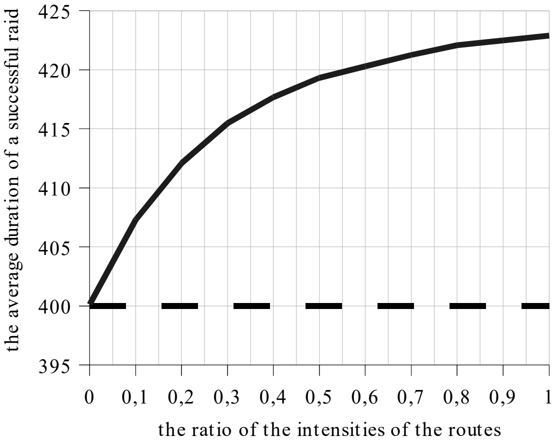


Fig. 10. The dependence of the average duration of a successful raid on the ratio of the intensities of the routes I_{ACE2DB}/I_{ACE1DB}, after 50,000 cycles.

Graphs of the dependence of the number of successful raids on the ratio of the intensities of the pheromone labels of the main and alternative routes I_{ACE2DB}/I_{ACE1DB} after 50,000 cycles of the original adapted ant algorithm and the oriented ant algorithm are shown in Fig. 11. The presented graphs show that, in the case of the oriented ant algorithm, with the growth of the I_{ACE2DB}/I_{ACE1DB} ratio, the number of effective raids does not decrease and is equal to the number of effective raids observed when there is only one route (I_{ACE2DB}/I_{ACE1DB} =0).

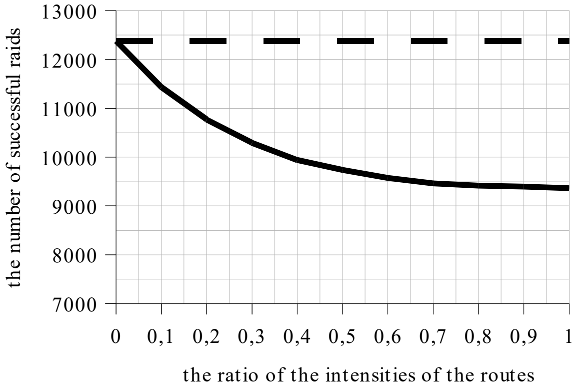


Fig. 11. The dependence of the number of successful raids on the ratio of the intensities of the routes I_{ACE2DB}/I_{ACE1DB}, after 50,000 cycles.

The obtained results prove that, in the case of using the oriented ant algorithm, robots moving through a maze of pheromone trails do not move in a circle in cycles formed by pheromone trails. At the same time, the oriented ant algorithm does not affect the robot's choice of alternative routes. That is, every time the robot finds itself at a vertex from which several alternative pheromone trails exit, the probability of its choosing one or another trail will be proportional to the intensity of its pheromone labels (pheromone trails entering this vertex are excluded from consideration).

CONCLUSIONS

Considering all of the above, the following conclusions can be drawn.

First, the presence of several alternative pheromone trails that have mutual intersections can cause a decrease in the efficiency of the adapted ant algorithm (i.e., an increase in the duration of the raid and a reduction in the number of successful raids) due to the possibility of robots moving in a circle in cycles formed by pheromone trails.

Secondly, in the case of the adapted ant algorithm, the consequences of the presence of several alternative pheromone trails that have mutual intersections for the duration of the raid and the number of successful raids depend on the ratio of the intensities of the pheromone labels of the main and alternative pheromone trails: the greater the ratio of the intensities of the pheromone labels, the longer raid duration and less number of productive raids.

Thirdly, the performance of the oriented ant algorithm was proven by computer simulation and it was shown that the effectiveness of this algorithm does not depend on the presence of several alternative pheromone trails that have mutual intersections.

REFERENCES

[1] Vardy, A. The swarm within the labyrinth: planar construction by a robot swarm. Artificial Life and Robotics. 28, 117–126 (2023). https://doi.org/10.1007/s10015-022-00849-5

[2] Alkilabi, M.H.M., Narayan, A. & Tuci, E. Cooperative object transport with a swarm of e-puck robots: robustness and scalability of evolved collective strategies. Swarm Intelligence. 11, 185–209 (2017). https://doi.org/10.1007/s11721-017-0135-8

[3] Morimoto, D., Hiraga, M., Shiozaki, N. et al. Evolving collective step-climbing behavior in multi-legged robotic swarm. Artificial Life and Robotics 27, 333–340 (2022). https://doi.org/10.1007/s10015-021-00725-8

[4] Ordaz-Rivas, E., Rodriguez-Liñan, A. & Torres-Treviño, L. Autonomous foraging with a pack of robots based on repulsion, attraction and influence. Autonomous Robots 45, 919–935 (2021). https://doi.org/10.1007/s10514-021-09994-5

[5] Hecker, J.P., Moses, M.E. Beyond pheromones: evolving error-tolerant, flexible, and scalable ant-inspired robot swarms. Swarm Intelligence 9, 43–70 (2015). https://doi.org/10.1007/s11721-015-0104-z

[6] Castello, E., Yamamoto, T., Libera, F.D. et al. Adaptive foraging for simulated and real robotic swarms: the dynamical response threshold approach. Swarm Intelligence. 10, 1–31 (2016). https://doi.org/10.1007/s11721-015-0117-7

[7] Adams, S., Jarne Ornia, D. & Mazo, M. A self-guided approach for navigation in a minimalistic foraging robotic swarm. Autonomous Robots. (2023). https://doi.org/10.1007/s10514-023-10102-y

[8] Talamali, M.S., Bose, T., Haire, M. et al. Sophisticated collective foraging with minimalist agents: a swarm robotics test. Swarm Intelligence. 14, 25–56 (2020). https://doi.org/10.1007/s11721-019-00176-9

[9] F. V. Abramov, “Implementation of an Adapted Ant Algorithm in the Presence of Substitute and Complementary Resources. Modeling the Behavior of the Manufacturer,” 2021 IEEE 2nd KhPI Week on Advanced Technology (KhPIWeek), 2021 - Conference Proceedings, pp. 341-345. doi.org/10.1109/KhPIWeek53812.2021.9570095.

Modified Genetic Algorithm with Enhanced Gene Correction for Optimal Class Scheduling in Higher Education Institutions

Oleh Zanevych
Department of applied mathematics
Ivan Franko National University of Lviv
Lviv, Ukraine
Oleh.Zanevych@gmail.com

Abstract — **In the modern educational landscape, ensuring efficient and conflict-free class scheduling in higher education institutions remains a paramount challenge. Traditional methods, although functional, often fall short in accommodating dynamic class environments and curriculums. This research introduces a unique genetic algorithm that addresses these challenges by targeting and eliminating undesirable genes during the crossover and mutation processes. Unlike conventional models that utilize a fitness function, our approach employs an objective function that takes smaller values for better scheduling configurations, ensuring a more precise evaluation. The algorithm is structured to prioritize the mutation of bad genes based on a localized objective function. When applied to higher education scheduling, the algorithm takes into account potential class overlaps, room allocation conflicts, and other common scheduling issues. Extensive numerical experiments demonstrate the effectiveness of this method, with results indicating the possibility of achieving a zero value for the objective function, representing a schedule devoid of any conflicts. This research not only offers a novel approach to class scheduling but also opens avenues for applying the algorithm to other complex scheduling scenarios.**

Keyword — genetic algorithm, crossover, mutation, objective function, scheduling, higher education, gene correction.

I. Introduction

Class scheduling in higher education resembles a complex jigsaw puzzle, with each piece representing classes, instructors, rooms, and time slots that must seamlessly fit together. As institutions expand and courses diversify, traditional scheduling methods, which are often manual or use basic algorithms, struggle to prevent issues like classroom shortages or overlapping classes. These inefficiencies disrupt academic routines and strain administration. Genetic algorithms (GAs), inspired by natural selection, have shown potential in addressing these optimization challenges, navigating vast solution spaces to find optimal outcomes. Although applied in various sectors, there's ample scope for refining GAs in class scheduling.

This research aims to bridge the gap between the potential of GAs and the practical challenges of class scheduling. We introduce optimized genetic algorithm that deviates from traditional models by placing an emphasis on the correction of undesirable genetic information. By proactively targeting and eliminating 'bad genes' during crucial operations like crossover and mutation, our approach ensures a more efficient evolution towards optimal schedules. Furthermore, our use of an objective function, as opposed to the commonly used fitness function, promises a more nuanced and accurate evaluation of scheduling solutions.

II. Background and Related Work

GAs, inspired by natural selection, were formalized in the 1960s by researchers like John Holland. Starting with a random set of potential solutions, GAs use selection, crossover, and mutation processes to optimize solutions. The effectiveness of each solution is assessed using a fitness function, and the process iterates until an optimal or satisfactory solution is found. They're used in various fields, from engineering to financial forecasting. In our study, we leverage GAs to address class scheduling challenges in higher education, aiming for conflict-free timetables.

Class scheduling has been a research focus for years, with GAs offering innovative solutions. Colorni, Dorigo, and Maniezzo explored high school timetabling using GAs to manage both fixed elements and teacher preferences in 1992 [1]. Abramson, in 1991, emphasized GAs' role in exam timetabling, particularly mutation operations [2]. In 2003, Burke, Petrovic, and Qu highlighted GAs' adaptability and integration with methods like simulated annealing for educational timetabling [3].

However, modern scheduling challenges demand new approaches. Current educational institutions face ongoing changes, requiring dynamic algorithms. Much existing research emphasizes hard constraints, often sidelining soft constraints vital for practical schedules. As institutions grow, scheduling becomes more complex, sometimes stretching GA efficiencies. A focused evolutionary approach, like our proposal targeting problematic genes, may enhance convergence and scheduling solutions [4-5].

Over the past decade, GAs have significantly evolved in higher education timetabling. This review highlights pivotal studies that have advanced GA methodologies for university and college class scheduling.

Wen-jing, W. [6] enhanced traditional AI-aided course scheduling, introducing an adaptive GA based on hard and soft constraints. This method surpassed conventional genetic algorithms in efficiency.

J. B. Matias et al. [7] addressed resource shortages with a hybrid genetic algorithm targeting both course scheduling and teaching workload, incorporating self-adaptive mechanisms and utilizing unused resources data structures.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

In the vocational education context at Airlangga University, Derawutie et al. [8] applied a Modified Genetic Algorithm (MGA) for optimizing scheduling across 21 diploma courses, effectively navigating constraints and ensuring efficiency.

Jing Xu & Zhihan Lv [9] developed an enhanced GA for college English course scheduling, incorporating a flexible decimal coding scheme and a local search operator for quicker convergence, outperforming traditional GAs in conflict resolution and fitness values.

Lastly, Zhang, Qiang [10] improved GAs using coevolution, catering to the complexities arising from expanding universities. Their algorithm's speed and optimal solutions emphasized its promise for contemporary course scheduling needs.

These studies highlight the progress in GAs for university scheduling. Through adaptive techniques, hybrid solutions, and course-specific modifications, the field displays ongoing potential for optimization amidst growing challenges. While foundational work remains strong, there's evident room for refinement to meet contemporary educational demands.

III. Method

For a proficient higher education class scheduling system, precise data input and representation are crucial. We outline the input data's structure and semantics, along with the class representation, providing a foundational framework for the genetic algorithm's functions.

A. Input Data for Timetabling

The essence of class scheduling relies on precise and thorough input data, sourced from institutional curricula and teaching loads. This data integrates requirements vital for organizing specific classes, which are fundamental to our timetabling application. Three primary tenets encompass these requirements:

- Lecturer: The faculty member assigned to conduct the class.
- Student Groups: An aggregated list detailing the specific student cohorts that are recipients of the teaching for the said class.
- Frequency of Classes: A quantitative measure indicating the number of times a particular class convenes in a week. It's imperative to note that this frequency is always in multiples of 0.5. To elucidate, a frequency of 1.5 suggests a bifurcated class schedule where the class meets once every week, and additionally, on a fortnightly basis, either in the first (numerator) or the second week (denominator).

B. Representation of Classes in the Schedule

With the foundational input data elucidated, the subsequent step entails the formulation of the initial population of schedules. This phase is pivotal as it sets the trajectory for the genetic algorithm's optimization processes. In this preliminary schedule population, each class— conceptualized as a gene within the genetic algorithm parlance — is depicted via a quintet of values:

- Requirement Index: Serving as a referential anchor, this index allows for quick retrieval of associated class details. For instance, through this index, one can discern the specific lecturer orchestrating the class or pinpoint the academic groups for which the class is being conducted.
- Day: Determines the particular day of the week the class is slated for.
- Time Slot Number: This signifies the chronological sequence of the class on the designated day.
- Location Index: Represents the spatial coordinates, indicating the specific classroom or venue where the class is to be conducted.
- Event Frequency: Categorically outlines whether the class is scheduled to occur on a weekly basis, or alternates between the first (numerator) or the second week (denominator).

C. Evaluation of Schedule Quality: Objective Function Incorporating Weighted Penalty Criteria

To ensure the optimization of the class scheduling system, we deploy a set of penalty criteria to evaluate the quality of a proposed schedule. These criteria allow for the identification and quantification of discrepancies and suboptimal configurations in the schedule, offering critical feedback to the genetic algorithm for further refinement.

Given:

- Lecturers are enumerated from 1 to L,
- Student groups span from 1 to G,
- Class venues are indexed from 1 to A.

For any arbitrary schedule configuration s, we define the following criteria of its incongruity:

- $\Pi_1(s,i)$ represents the count of time overlaps or collisions pertaining to the lecturer with the index i. Time overlaps occur when a lecturer is slated to conduct classes at two different venues simultaneously. A higher value indicates a greater frequency of such overlaps, underscoring a pressing need for schedule revision for the concerned lecturer.
- $\Pi_2(s,j)$ denotes the count of temporal overlaps within student groups indexed by j. Such overlaps are indicative of situations where a particular student group is expected to attend multiple classes concurrently, highlighting a glaring incompatibility in the schedule.
- $\Pi_3(s,t)$ symbolizes the frequency of scheduling overlaps at a specific class venue indexed by t. A venue overlap suggests that more than one class has been allocated to the same venue at the same time slot, necessitating immediate rectification.
- $\Pi_4(s,k)$ specifies the number of idle periods or "windows" experienced by the lecturer indexed by k. While some windows might be intentional, providing lecturers respite between classes, excessive windows can disrupt the teaching flow and lead to inefficient utilization of teaching resources.
- $\Pi_5(s,p)$ indicates the frequency of windows within the schedule of student groups indexed by p. Like with lecturers, while a minimal number of breaks can be beneficial, overextension of idle periods can result

in suboptimal learning experiences and prolonged academic days.

To effectively gauge the incompatibility of a proposed schedule **s**, an objective function, *o.f.(s)*, is crafted to cumulatively account for the individual incongruities while weighing them in accordance with their severity and implications. The objective function is defined as:

$$o.f.(s) = \beta_1 \sum_{i=1}^{L} \Pi_1^{\alpha_1}(s,i) + \beta_2 \sum_{j=1}^{G} \Pi_2^{\alpha_2}(s,j) + \beta_3 \sum_{t=1}^{A} \Pi_3^{\alpha_3}(s,t) + \beta_4 \sum_{k=1}^{L} \Pi_4^{\alpha_4}(s,k) + \beta_5 \sum_{p=1}^{G} \Pi_5^{\alpha_5}(s,p) \quad (1)$$

Function (1) incorporates two pivotal components that reflect the gravity of each incompatibility:

- **Weightage Parameters (β_i).** These parameters assign a relative importance to the different penalty criteria. Given that $\beta_1, \beta_2, \beta_3 \gg \beta_4, \beta_5 > 0$, it underscores that time overlaps concerning lecturers, student groups, and class venues are deemed significantly more detrimental than the occurrence of windows in schedules. This prioritization resonates with the tangible consequences of these incongruities — while overlaps directly impede the teaching-learning process, windows, though suboptimal, do not render the schedule unviable.

- **Exponential Parameters (α_i).** With $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5 > 1$, the exponential parameters magnify the penalties in cases of recurring overlaps or windows for specific entities. This is predicated on the understanding that isolated instances of discrepancies might be rectifiable or tolerable, but repetitive errors for the same entity (be it lecturer, student group, or venue) amplify the challenges and thus warrant escalated penalties.

The proposed objective function stands out in its ability to offer a nuanced, differential assessment of schedule quality. Instead of treating each incongruity with uniform severity, it meticulously discerns between the nature and recurrence of discrepancies.

- **Granular Assessment**. By adjusting the β_i and α_i parameters, the function provides the flexibility to calibrate the penalties based on institutional priorities and constraints.

- **Reinforcing Robustness**. The exponential augmentation ensures that schedules with recurrent issues for specific entities are penalized more heavily. This nudges the genetic algorithm towards crafting schedules that are not just free of overlaps and windows, but also balanced and equitable in their distribution of classes.

- **Operational Practicality**. The weighted approach aligns with real-world scheduling challenges. In practice, a schedule with occasional windows might be acceptable, but one with regular class overlaps is operationally infeasible.

In the detailed structure of our scheduling objective function (1) it's crucial to highlight a nuanced aspect: the values $\Pi_1(s,i)$, $\Pi_2(s,j)$, $\Pi_3(s,t)$, $\Pi_4(s,k)$ and $\Pi_5(s,p)$ are not strictly integers but can be multiples of 0.5. This fractional representation is not an arbitrary choice but embodies the unique scenario of bi-weekly classes, held either in the "numerator" or the "denominator" pattern. Such classes may experience restrictions — like time overlays or windows — once every two weeks, leading to these half-unit deviations. By incorporating this granularity, the objective function offers a refined and precise evaluation mechanism, ensuring the algorithm remains attuned to both regular and sporadic scheduling challenges.

In summary, the proposed objective function, with its weighted and exponential penalty system, offers a comprehensive, pragmatic, and adaptable framework to evaluate and optimize class schedules, ensuring alignment with both academic objectives and operational realities.

D. Evaluating Gene Quality with the Local Objective Function

In the intricate endeavor of class scheduling optimization, the utility of global objective functions can sometimes overshadow the necessity of nuanced, gene-specific evaluation. To address this, we introduce a specialized measure termed the local objective function, designed to gauge the perturbations instigated by a singular class or "gene" within the broader genetic algorithm framework.

For any given schedule variant *s* and a specific class index *i*, the local objective function *l.o.f.(s,i)* is defined as:

$$l.o.f.(s) = \beta_1 \Psi_1^{\alpha_1}(s,i,l) + \beta_2 \sum_{j \in g} \Psi_2^{\alpha_2}(s,i,j) + \beta_3 \Psi_3^{\alpha_3}(s,i,d) + \beta_4 \Psi_4^{\alpha_4}(s,i,l) + \beta_5 \sum_{p \in g} \Psi_5^{\alpha_5}(s,i,p) \quad (2)$$

Here, *l* represents the index of the lecturer conducting the class, *d* signifies the index of the classroom, and *g* encompasses the set of student groups attending the class. This function meticulously evaluates disruptions:

- $\Psi_1(s,i,l)$ denotes time overlaps of the lecturer with index *l* during class *i*.

- $\Psi_2(s,i,j)$ measures the time overlaps experienced by the student group with index *j* during class *i*.

- $\Psi_3(s,i,d)$ captures overlaps at the location indexed by *d* during class *i*.

- $\Psi_4(s,i,l)$ signifies the windows or gaps encountered by the lecturer *l* on the day of class *i*.

- $\Psi_5(s,i,p)$ represents the windows experienced by the student group *p* on the day class *i* is conducted.

In subsequent iterations, (2) proves indispensable for pinpointing "bad genes". By quantifying the discrepancies introduced by each class in the schedule, the function lays the groundwork for targeted improvements, either through gene replacement during crossovers or strategic mutations, ensuring the genetic algorithm consistently converges towards optimal solutions.

E. Crossover Procedure: replace "bad" genes

GAs excel by merging strengths of solutions and discarding weak components. Central to this is the crossover procedure. In our model, it plays a vital role, focusing on replacing weaker genes with stronger ones from another schedule, leading to a more optimized result.

In our approach, the initial phase involves a random selection of two distinct schedule variants, denoted s_1 and s_2, from the overarching population *P*. To understand the magnitude of our gene pool, let V represent the total number of classes, equivalently referred to as genes, in any given schedule.

The cornerstone of our method is the determination of how many genes will participate in the crossover. To this end, we generate a random number r constrained by the formulae $CR_{min} \times V \leq r \leq CR_{max} \times V$, where both CR_{min} and CR_{max} are predefined constants from the interval (0,1). It's paramount to note the inherent boundaries set for these constants: $1 \leq CR_{min} \times V < CR_{max} \times V \leq V - 1$. This ensures a balanced and regulated transfer of genetic material.

The innovation in our crossover mechanism lies in its strategy to select the genes for exchange. Rather than relying on random or fixed position-based crossovers, we specifically target genes that are "sub-optimal" or "bad" in schedule s_1. The metric to gauge this is none other than the local objective function (2). We earmark the top *r* genes with the highest *l.o.f.* values in s_1 and substitute them with their counterparts from s_2, thereby crafting a novel schedule variant s'.

To robustly navigate through the solution space, this crossover procedure is reiterated $\lceil C \times S \rceil$ times, where C is a prescribed constant from the interval *(0,1)* and S — the number of schedule variants in the population P. Such consistent application ensures that the offspring schedules are routinely infused with superior genetic material, driving the population towards enhanced optimization with every generation.

In essence, our crossover methodology is both selective and adaptive, built on the foundation of recognizing weaknesses in one schedule and actively seeking to mitigate them with strengths from another.

F. Mutation Procedure: Refining Genes for Optimized Scheduling

In the optimization process within GAs, mutation plays a crucial role in diversifying the gene pool and facilitating the escape from local optima. Our focus on the mutation procedure centers around the amelioration of suboptimal genes in the schedule, a procedure pivotal to refining solutions.

For any given schedule variant *s* within the population *P*, we administer mutations to a subset of classes characterized by the highest values of the local objective function (2). This subset comprises $\lceil M \times V \rceil$ classes, where M is a predetermined fraction and V signifies the total number of classes or genes. The purpose of the mutation operation is to refine these genes, possibly modifying several of their attributes — be it the day, time slot, location index, or the frequency for bi-weekly classes.

Elaborating on the methodology for mutating the day of a specific class: a random number r_d is generated within the range [0, 1]. When r_d is less than or equal to a predefined threshold MR_d , the day value for that class is then randomized. Analogous procedures are followed for the mutation of the time slot, location index, and lesson frequency using respective randomly generated numbers r_t, r_l, and r_f, each compared against their specific thresholds MR_t , MR_l and MR_f.

An important distinction is made concerning the frequency mutation: it is exclusively applied to classes that operate on a bi-weekly basis — either in the numerator or denominator. Classes that occur weekly remain unaffected in terms of frequency.

Concludingly, the mutation procedure does not merely adjust the existing schedule but generates a novel variant that enriches the population. The original version is retained, ensuring that the mutation operation does not directly alter it, but rather contributes a refined variant to the pool. This method underscores our strategy to evolve the population progressively while preserving diversity and mitigating the risk of stagnation.

G. GA for Schedule Generation: A Comprehensive Overview

In the ever-evolving field of schedule optimization, the utilization of GAs offers unparalleled potential. The outlined algorithm presents an advanced approach to generating class schedules for institutions of higher education, balancing efficacy with constraint adherence.

- **Initialization**. Initiated by a set of predefined requirements, the algorithm commences by randomly crafting an initial population comprising of *S* schedule variants, where *S* is a predetermined constant.
- **Crossover Operation**. The intricacy of the crossover mechanism stands highlighted in its ability to infuse fresh genetic material into the existing population. By implementing this procedure, the population augments by an additional $\lceil C \times S \rceil$ schedule variants, fostering diversity and enhancing the explorative capacity of the algorithm.
- **Mutation Operation.** Subsequent to the crossover, the mutation procedure is invoked. Its quintessential purpose lies in refining schedules by tweaking genes identified as suboptimal. Remarkably, this procedure duplicates the population size as for every schedule variant, a modified counterpart is generated, targeting the rectification of unfavorable genes.
- **Post-operative population dynamics**. Consequent to the aforesaid operations, the population burgeons to encompass $2 \times \lceil 1 + C \rceil \times S$ schedule variants. This augmented set signifies a mosaic of original, crossed, and mutated schedules.
- **Selection Procedure**. To streamline this expanded population, a rigorous selection procedure is employed. It meticulously cherry-picks *S* schedule variants that boast the most minimal values of the objective function (1). This ensures that the subsequent iterations of the algorithm operate on a refined subset of promising schedules.
- **Iterative Convergence**. The algorithm repeatedly invokes the trinity of crossover, mutation, and selection procedures in alternation. The cessation criteria for this iterative process are twofold: the attainment of a pristine schedule variant with an objective function value of zero, symbolizing a schedule devoid of any constraint violations; or, in cases where such an ideal is elusive, the completion of a stipulated maximum number of iterations *K*.

- **Key Assurances**. A hallmark of this genetic algorithm iteration is its robustness against discrepancies in academic hours. Initial schedules are diligently curated to enshrine the precise number of academic hours for each class. Moreover, the subsequent crossover and mutation procedures are architected to be conservative regarding this attribute, ensuring that schedules never falter in terms of class durations. Thus, concerns about classes being underrepresented or excessively scheduled are meticulously obviated.

IV. Results and Discussion

To evaluate the efficacy of the proposed modified genetic algorithm, a software application was developed using Java 20. For the purposes of testing, input data were randomly generated to create class schedules of varying dimensions across three distinct experiments.

A. Parameter Settings

- **Objective function parameters**. The weightage values were selected as $\beta_1 = 150$, $\beta_2 = 100$, $\beta_3 = 50$, $\beta_4 = 5$ and $\beta_5 = 20$. Concurrently, the exponential parameters across all five categories were set to a uniform value: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 2$.
- **The population size**, pivotal for the genetic algorithm's exploratory capabilities, was locked at S=20.
- **Crossover Dynamics**. For the crossover operation, the parameters were initialized as $C = 0.5$, $CR_{min} = 0.05$, and $CR_{max} = 0.2$, optimizing the balance between exploration and exploitation.
- **Mutation Specifications**. In the mutation operation, the general mutation rate parameter was set to $M = 0.1$, while the specific mutation ratios for day, time slot, location index, and frequency were uniformly set at $\mathrm{MR_d} = \mathrm{MR_t} = \mathrm{MR_l} = \mathrm{MR_f} = 0.25$.

B. Experimental Design

To subject our model to a range of complexities and demands, three disparate datasets were crafted, each varying in its internal structural parameters:

- **Experiment 1.** Incorporated a setting of $V = 200$ classes, $L = 10$ lecturers, $G = 5$ student groups, and $A = 10$ class venues.
- **Experiment 2.** Elevated the dimensions to $V = 400$ classes, $L = 20$ lecturers, $G = 10$ student groups, and $A = 20$ class venues.
- **Experiment 3.** Positioned at the highest complexity level with $V = 800$ classes, $L = 40$ lecturers, $G = 20$ student groups, and $A = 40$ class venues.

Figures 1 to 3 depict the declining value of the objective function (1) for each iteration, illustrating the algorithm's journey towards optimization.

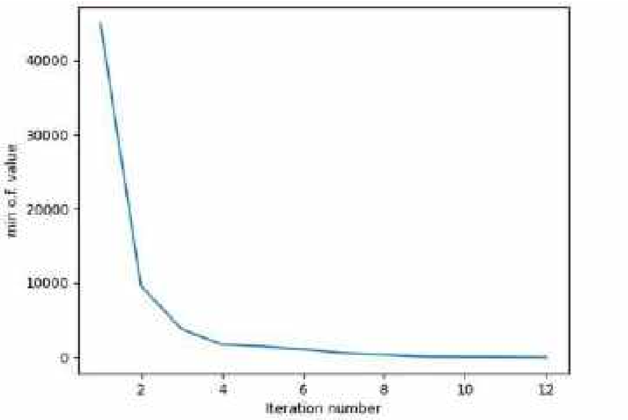


Fig. 1. Experiment 1. V=200, L=10, G=5, and A=10.

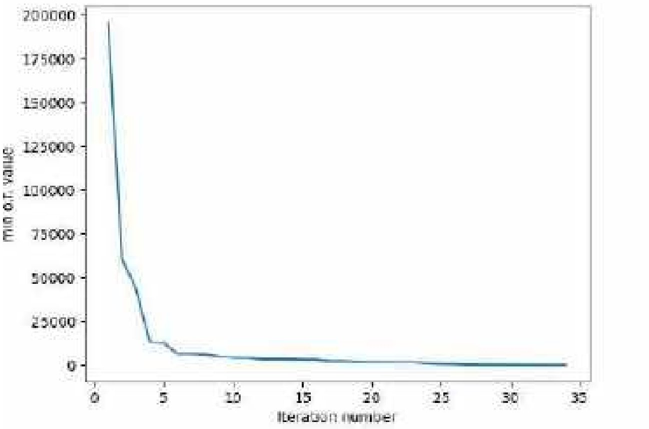


Fig. 2. Experiment 2. V=400, L=20, G=10, and A=20

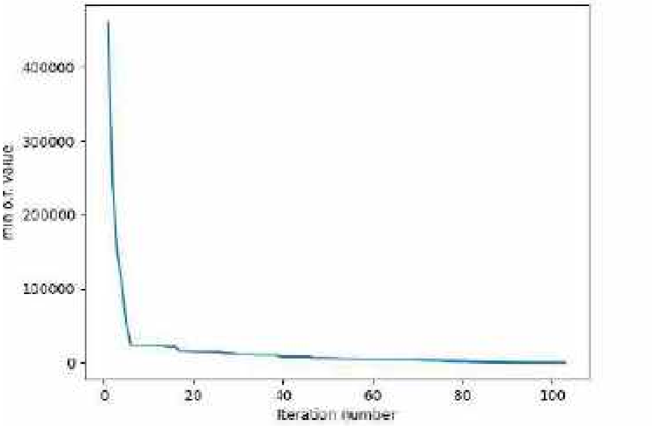


Fig. 3. Experiment 3. V=800, L=40, G=20, and A=40

In experiments with varying complexities, the algorithm required 12, 34, and 103 iterations for Experiments 1, 2, and 3 respectively to achieve this ideal result (the objective function (1) successfully converged to zero).

In recent computational experiments, significant emphasis was placed on the merits of selectively replacing bad genes during the crossover operation and specifically modifying these adverse genes in the mutation phase. This approach was juxtaposed against a conventional method of the GA, wherein the crossover operation involved a general exchange of genes and the mutation phase entailed arbitrary gene modifications, rather than a targeted alteration of the unfavorable ones.

Evaluating the efficiency of both algorithms, three distinct sets of input data were processed. The results manifested noticeable disparities in the performance of the two methodologies. When the conventional genetic algorithm was applied to the data from the initial experiment, the value of the objective function only touched 200 post the 490th iteration. Following this, there was a notable stagnation in the optimization rate, culminating in the objective function attaining zero only after a protracted 6,343 iterations. This progression can be visually discerned in Figure 4.

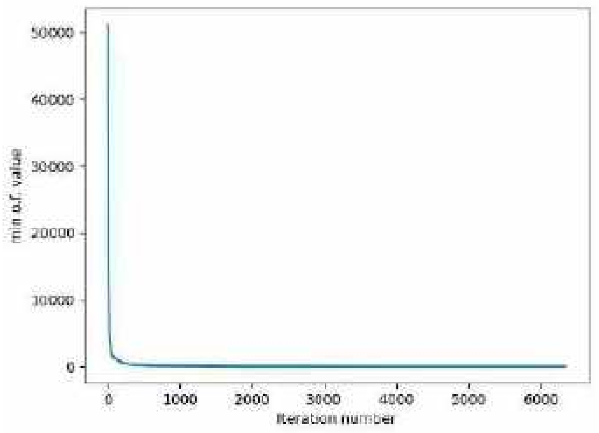


Fig. 4. Traditional GA Performance: Experiment 1 Data

On the contrary, for the data derived from the subsequent experiments (2 and 3), the conventional version of the GA struggled significantly. Notably, even after a considerable 10,000 iterations, it was unable to achieve a zero value for the objective function (1). The objective function values stagnated at 2,455 and 17,865 for the data extracted from experiments 2 and 3, respectively. Such figures underscore a significant deviation, suggesting that the traditional genetic algorithm may be less adept at minimizing errors when compared to its modified counterpart.

Conclusions

Our modified genetic algorithm introduces a novel approach by integrating a local objective function to accurately identify and address "bad genes." This innovative technique emphasizes replacing these genes during crossover and mutation operations. This targeted strategy significantly accelerates the algorithm, especially beyond a certain iteration threshold where traditional GAs tend to decelerate considerably. In all three experiments, our method showcased a remarkable speed-up of over 1000 times compared to the conventional GA variant.

By ensuring a precise crossover and mutation strategy that prioritizes the preservation of academic hour integrity, the algorithm demonstrates both efficacy and efficiency. Tested across diverse complexities, it consistently generated constraint-free schedules promptly, attesting to its profound potential in addressing contemporary educational scheduling dilemmas.

Acknowledgment

We extend our profound gratitude and respect to the valiant armed forces of Ukraine, who stand as the unwavering bulwark against the full-scale invasion by the russian federation. Their indomitable spirit, dedication, and sacrifices in the defense of our homeland, sovereignty, and the democratic values that Ukraine holds dear are truly commendable.

References

[1] Colorni, Alberto & Dorigo, Marco & Maniezzo, Vittorio. (1994). A Genetic Algorithm To Solve The Timetable Problem.

[2] D. Abramson, (1991) Constructing School Timetables Using Simulated Annealing: Sequential and Parallel Algorithms. Management Science 37(1):98-113. DOI: 10.1287/mnsc.37.1.98.

[3] Burke, Edmund & Petrovic, Sanja & Qu, Rong. (2006). Case Based Heuristic Selection for Timetabling Problems. Journal of Scheduling. 9. 115-132. DOI: 10.1007/s10951-006-6775-y.

[4] Pillay, N. A survey of school timetabling research. Ann Oper Res 218, 261–293 (2014). DOI: 10.1007/s10479-013-1321-8.

[5] Pillay, N. (2013). A survey of school timetabling research. Annals of Operations Research. DOI: 218. 10.1007/s10479-013-1321-8.

[6] Wen-jing, W. (2018). Improved Adaptive Genetic Algorithm for Course Scheduling in Colleges and Universities. International Journal of Emerging Technologies in Learning (iJET), 13(06), pp. 29–42. DOI: 10.3991/ijet.v13i06.8442.

[7] J. B. Matias, A. C. Fajardo and R. P. Medina, "A Hybrid Genetic Algorithm for Course Scheduling and Teaching Workload Management," 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM), Baguio City, Philippines, 2018, pp. 1-6. DOI: 10.1109/HNICEM.2018.8666332.

[8] Derawutie, D., Wuryanto, E., & Jie, F. (2018). Course scheduling using modified genetic algorithm in vocational education. International Journal of Operations and Quantitative Management, 24(3), 203-210.

[9] Jing Xu & Zhihan Lv, 2021. "Improved Genetic Algorithm to Solve the Scheduling Problem of College English Courses," Complexity, Hindawi, vol. 2021, pages 1-11, June. DOI: 10.1155/2021/7252719.

[10] Zhang, Qiang. "An optimized solution to the course scheduling problem in universities under an improved genetic algorithm" Journal of Intelligent Systems, vol. 31, no. 1, 2022, pp. 1065-1073. DOI: 10.1515/jisys-2022-0114.

Processing Sensor Signal Under Low Values of Signal to Noise Ratio

Zinovii Liubun
Department of RadioPhysic and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
zinovijlyubun@gmail.com

Bohdan Bryk
Infineon Technologies
Lviv, Ukraine
Bohdan.Bryk@infineon.com

Vasyl Mandziy
Infineon Technologies
Lviv, Ukraine
Vasyl.Mandziy@infineon.com

Oleksandr Karpin
Infineon Technologies
Lviv, Ukraine
Oleksandr.Karpin@infineon.com

Bogdana Kalivoshka
Department of RadioPhysic and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
bogdana.kalivoshka@lnu.edu.ua

Serhiy Velhosh
Department of RadioPhysic and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
serhiy.velhosh@lnu.edu.ua

***Abstract* — Creating effective means for the implementation of the filtering algorithms is an important task. The method significantly reduces the cost of the implementation due to the unnecessity of involving high-level experts and creating adaptive means, whereas it ensures the maximum speed with minimum computing resources. Such processing algorithms are expected to provide the possibility of easy implementation on microcontrollers and microprocessors.**

The usage of the neural network (NN) approach provides the possibility to quickly and with insignificant costs implement the compact NNs for filtering noise. This article considers three structures of NNs for filtering noisy signals.

Herein, you can learn the following about NNs:

- **the possibility of their fast adaptation – tuning or training**
- **the proof of their effectiveness for the emittance of signal under low value of signal to noise ratio**
- **the results, which confirm the possibility to obtain simple NNs for signal filtering under high noise levels.**

Keywords — neural network, low signal to noise ratio

I. Introduction

The usage of NNs of various structures allows obtaining adaptive systems of data analysis [1-10]. A simplest NN whose structure fully corresponds to the recurrent digital IIR filter (Fig. 1) is described with the equation:

$$Out_k = \sum_{i=0}^{m-1} Wx_i \cdot x_{k-i} + \sum_{i=0}^{n-1} Wy_i \cdot x_{k-(i+1)} \qquad (1)$$

The results of training the NN, which corresponds to the structure of the recurrent digital filter are given in [9].

It is known that while implementing such a type of filters, certain correlations between the coefficients are required to ensure their stability. For example, under $m = n = 1$, it is necessary to meet the condition $Wx_0 + Wy_0 = 1$. While training a NN, it may happen that the obtained values of the scales will not comply with the requirements. For various compositions of training sets, filters with different coefficients are obtained. If the filter coefficient values do not meet this requirement, the point may be that the selection of the training set of signal was wrong. Usually, in this case, the filtering still occurs but the signal level reproduction is poor. Then, to obtain better results, the correction of the training sets is required involving all possible types of noisy signals.

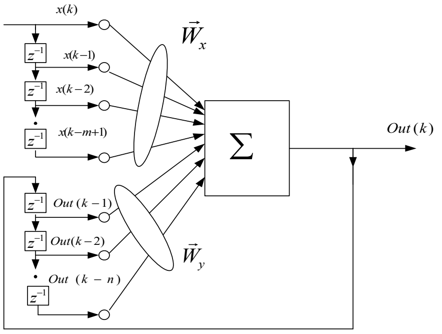


Fig. 1. Linear one-neuron recurrent NN

The hope is that results could be improved when using more complicated NNs. Complication of the NN structure will require the transition to multi-layer NNs. Apparently, in this case, non-linear functions of activation are required, otherwise, the multi-layer NN with the linear function of activation is equivalent to the one-layer NN. Clearly, increasing the number of weight values and the non-linear activation function will require more computational resources. As a first step towards transitioning to a multi-layer neural network with a non-linear activation function, a two-layer neural network with the ReLU activation function shown in Fig. 2 is proposed.

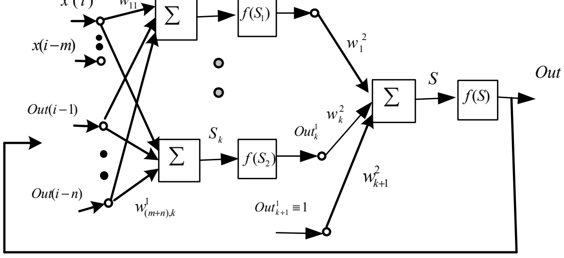


Fig. 2. Structure of two-layer NN

The next step is using the sigmoidal function of activation for the similar NN structure. The training of NNs was executed with the gradient method with the adaptive training speed. While implementing digital filters through the NN

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

training, the selection of the training set is very important. To train NNs, data simulated under various values of signal to noise ratio and touch duration was used.

To estimate the filter effectiveness, such criteria were used

- reducing the noise level
- delayed reaction of the filter to appearance/disappearance of the signal
- reproduction of the level of the useful signal.

While working with sensor panels, most important are the following factors: detection of the touch itself and its duration under low signal to noise ratio.

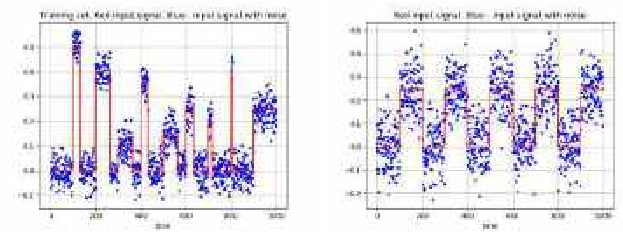
II. Results

As it was mentioned earlier, an important task is the selection of a training set that can train the NN to recognize the required behavior:

- for most of possible cases of incoming signals
- the limit values of the parameters
    - the minimal signal to noise ratio
    - the minimal touch duration.

Digital experiments showed that the properties of filters that were obtained through training a NN significantly depend on the training set. Therefore, the training was executed with two sets of data:

- set SL1 – contains a set of signals of various levels and duration touch (Fig. 3 a)
- set SL2 – contains a fixed value of the level and duration of touch (Fig. 3 b).



(a) (b)

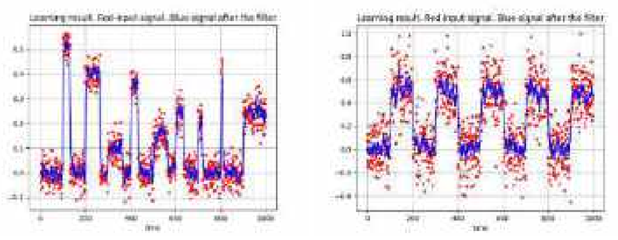
Fig. 3. a) Training signal SL1 under different levels and duration (SNR=2) b) – training signal SL2 under identical levels and duration (SNR=1)

The signals levels consider the limitation of the output of the sigmoidal function of the NN activation. The signals were mixed with white noise of the zero mean value and given dispersion. The correlation signal/noise was defined as follows:

$$SNR = \frac{h}{6 \cdot \sigma_{input}} \qquad (2)$$

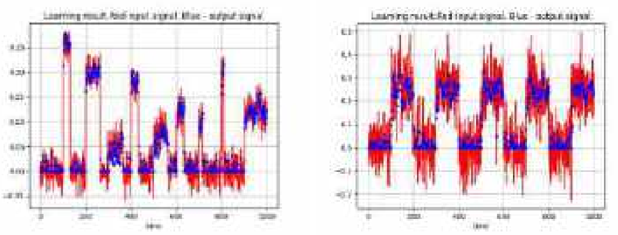
where h – the signal level, σ_{input}– standard deviation of the white noise.

The results of training NNs under the specified types of the training signal are showed in Fig. 4-6. The research was done for two-layer NNs with 16 neurons in the input layer and 1 neuron in the output layer. The research proved that such a structure can be considered the optimal. The reduction of the quantity of neurons in the input layer significantly deteriorates the results, whereas the increase of the quantity of neurons does not improve the results significantly. Also, the research proved that for training, noise levels higher than the limit levels of signals, which the given filter is expected to provide must be selected.



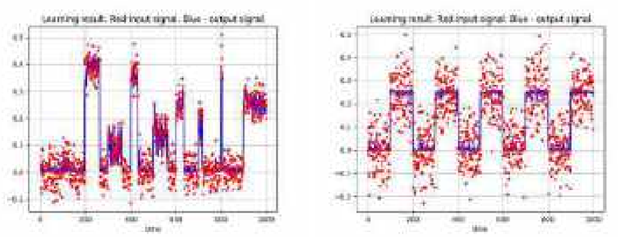
(a) (b)

Fig. 4. Linear filter: a) signal SL1; b) SL2



(a) (b)

Fig. 5. Two-layer NN with activation function ReLu: a) signal SL1; b) signal SL2



(a) (b)

Fig. 6. Two-layer NN with sigmoidal activation function: a) signal SL1; b) signal SL2

The main criterion to access the training quality was the reduction of the noise level after transition of the signal through the filter:

$$K = \frac{\sigma_{input}}{\sigma_{output}} = \frac{SNR_{output}}{SNR_{input}} \qquad (3)$$

The results of training are given in Table I.

Table I. The results of training NNs under two types of the training signals

#	NN	Type of signal	Average error of training	K
1	Linear filter	SL1	0.05	1.3
2		SL2	0.05	2.4
3	ReLu	SL1	0.02	1.4
4		SL2	0.03	2.2
5	Two-layer sigmoid	SL1	0.02	1.6
6		SL2	0.0075	12.3

Analyzing the data from Table I, the following can be concluded:

- The results of training and filters obtained after training significantly depend on the training signal due to the fact that the NN must be trained to recognize what we want to reproduce (do) first of all – reduce the noise or reproduce the spike and the level of the signal. Apparently, the requirements must be balanced.

- When training NNs with signal SL1, the average error for different filters does not differ significantly as well as the value of the noise reduction K. Complication of the NN under this signal is not effective. The selection of the training signal was wrong.
- Under SL2 training signal, the training results are better and depend on the complexity of the NN. This is because a simpler or better-defined training signal indicates to the network what we want to achieve at the output.
- The nearly ideal result is obtained when training the NN with the sigmoidal activation function and structure 16-1 with SL2.

The difference in the filtering quality can also be explained in the following way:

Low-frequency filters are implemented with the help of NNs, but signal SL1 has a broad spectrum (availability of high-frequency components). So, the filter will either properly filter the noise and spoil the impulse reproduction or will poorly filter the noise. Under SL2 similar signal, the spectrum is narrower, therefore it is easier to separate the useful signal from noise.

Apparently, the conclusions from the training results must be fully confirmed by the testing of trained NNs.

All the filters were tested using the signal showed in Fig. 7 under different levels of noise (different values of SNR).

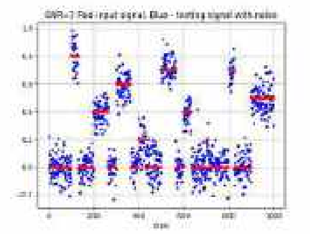


Fig. 7. Example of signal for testing (SNR=2)

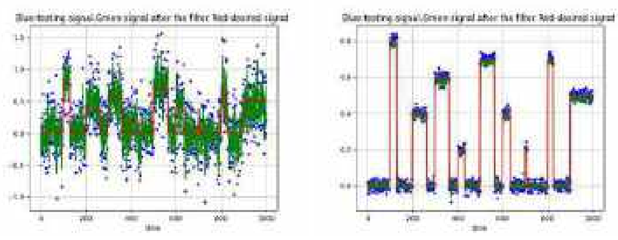
The results of testing under different levels of the input signal noise for the linear filter are given in Table II.

TABLE II. RESULTS OF TESTING UNDER DIFFERENT LEVELS OF INPUT SIGNAL NOISE FOR LINEAR FILTER

Linear filter						
Signal for training	Training signal SNR					
	0.5	1	2	4	8	16
SL1	-	1.5	1.5	1.4	1.5	1.6
SL2	3.5	3.7	3.4	3.1	4	3.6

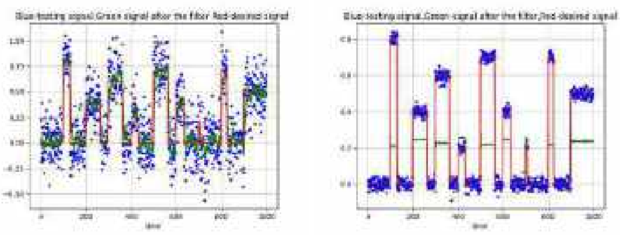
For the case of training signal SL1, under SNR < 1 values, the filtering does not occur and the signal at the output even does not respond properly to the input signal (Fig. 8 a). Under lower levels of noise, the filter works but the quality of the filtering is poor (Fig. 8 b). This can be explained by the wrong selection of the training signal.

Much better results are obtained with training signal SL2. Even under SNR < 2, the noise filtering works (Fig. 9 a) and this property remains under lower levels of the input noise signal.



a) SNR = 0.5 b) SNR = 8

Fig. 8. Test results for training signal SL1



a) SNR = 1 b) SNR = 8

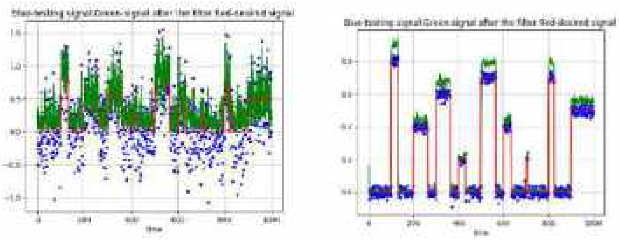
Fig. 9. Test results for training signal SL2

The results of testing are given in Table III: under different levels of the input signal noise, for the filter implemented with a two-layer NN with 16 neurons in the input layer, under the activation function ReLu.

TABLE III. RESULTS OF TESTING UNDER DIFFERENT LEVELS OF NOISE FOR ACTIVATION FUNCTION RELU

ReLu						
Signal for training	Training signal SNR					
	0.5	1	2	4	8	16
SL1	-	-	1.7	1.7	1.9	2.0
SL2	2.0	2.0	3.3	3.7	6.3	1.6

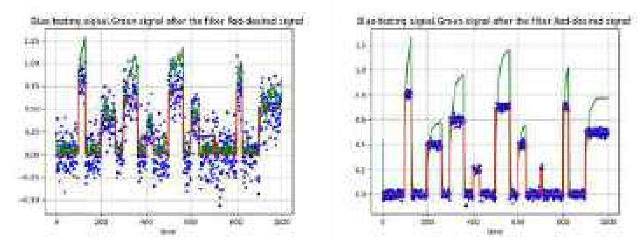
For training signal SL1 under SNR = 0,5 and SNR = 1 of the input signal, the signal on the output is not reproduced, so, no point talking about the filtering (Fig. 10 a). Under higher SNR values, the signal level is reproduced almost well although there is practically no filtered signal shift. But under such conditions, the noise reduction is insignificant, so, the usage of such conditions is pointless (Fig. 10 b).



a) SNR = 0.5 b) SNR = 8

Fig. 10. Test results for training signal SL1

For filters obtained under SL2, the situation is much better. The examples of the filter performance are showed in Fig. 11. Under SNR = 0,5 and SNR = 1, the signal is reproduced qualitatively. The filtration is present but the signal spike shape and level differ much from the actual signal. The signal level is reproduced only qualitatively. The flat surface of the signal is not reproduced. The time shift is insignificant. The filter can be used when only the time of the detection and availability of a touch are important (Fig. 11 b). Here, at SNR=16, the filtration turned out to be much smaller due to the peculiar shape of the impulse peak obtained as a result of the filtration – Fig. 11 b.



a) SNR = 1 b) SNR = 8

Fig. 11. Test results for training signal SL2

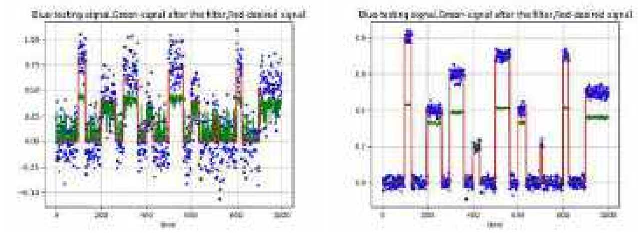
Apparently, in this case, only the noise filtering is important, not the defining of the signal level (value).

The final case is the results obtained when using a two-layer NN with the sigmoidal activation function. The results of the testing under different noise levels of the input signal for a filter implemented with the two-layer NN with 16 neurons in the input layer are given in Table IV.

TABLE IV. RESULTS OF TESTING UNDER DIFFERENT LEVELS OF NOISE FOR ACTIVATION FUNCTION RELU

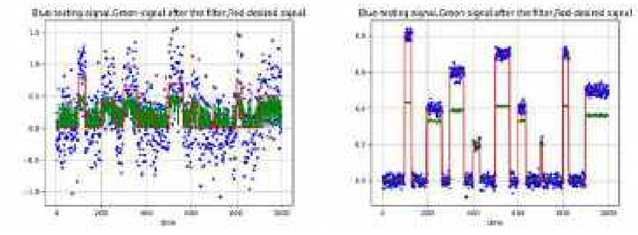
Sigmoidal function						
Signal for training	Training signal SNR					
	0.5	1	2	4	8	16
SL1	-	-	3.5	3.6	3.6	3.6
SL2	-	-	7.7	7.8	6.5	6.8

Under the SNR < 2 values, the signal is not reproduced (Fig. 12 a). Starting from the noise level SNR = 2, signal is reproduced and filtered.



a) SNR = 0.5 b) SNR = 8

Fig. 12. Test results for training signal SL1



a) SNR = 0.5 b) SNR = 8

Fig. 13. Test results for training signal SL2

Similarly to training signal SL1, under SL2, signal is not reproduced when SNR < 2 (Fig. 13 a). Starting from noise level under SNR = 2, signal is reproduced and filtered. The signal level is not reproduced, is stable, and a little smaller than the signal under which the NN was trained.

For this case, the K value is the biggest, which means that the filtering of high-frequency noise is the best. Another positive factor is that the delay of the signal front is practically absent and is 1-2 steps of discretization.

CONCLUSIONS

Apparently, the selection of the training set is very important for obtaining the best filters. For our case, it is obvious that the training set SL2 is better. Analyzing the results of the study of filters implemented using three types of neural networks allows us to make the following conclusions:

- The linear filter is the best if the filtering objective is simultaneous noise filtering and obtaining the signal level value under the minimal time shift. The linear filter: reproduces the signal level properly, gives a small-time shift, has the filtering properties – K = 3-4.
- The winner is the two-layer NN with sigmoidal function if the main task is to obtain the moment of touch appearance and disappearance, and determining the absolute level value is not necessary. The biggest value K = 6-8 under the minimal time shift.

REFERENCES

[1] Geoff Walker, 2020. [Online]. Available: https://www.walkermobile.com/SID_2014_ID_Touch_Technology_Review.pdf.

[2] Multi-Touch – Part 2: Filtering and Touch Detection, 2020. [Online]. Available: https://larrylisky.com/2014/03/11/multi-touch-part-2-filtering-and-touch-detection/; 2020.

[3] S. Haykin. Kalman Filtering and Neural Networks. First published:1 October 2001. Print ISBN:9780471369981 Online ISBN:9780471221548 DOI:10.1002/0471221546

[4] Bryan Lim, Stefan Zohren and Stephen Roberts. Recurrent Neural Filters: Learning Independent Bayesian Filtering Steps for Time Series Prediction Oxford-Man Institute of Quantitative Finance Department of Engineering Science University of Oxford Oxford, UK {blim,zohren,sjrob}@robots.ox.ac.u.

[5] Oxford-Man. Recurrent Neural Filters: Learning Independent Bayesian Filtering Steps for Time Series Prediction. Retrieved from https://www.oxford-man.ox.ac.uk/wp-content/uploads/2020/03/Recurrent-Neural-Filters-Learning-Independent-Bayesian-Filtering-Steps-for-Time-Series-Prediction.pdf

[6] IOPSCIENCE. Recurrent neural networks as approximators of non-linear filters operators. Retrieved from https://iopscience.iop.org/article/10.1088/1742-596/1141/1/012115/pdf

[7] Stowers Institute Research Websites. A fast noise filtering algorithm for time series prediction using recurrent neural networks. Retrieved from https://research.stowers.org/bru/RNN_Filter2_V3.pdf

[8] A.G. Parlos, S.K. Menon, A.F. Atiya. An algorithmic approach to adaptive state filtering using recurrent neural networks. “IEEE Transactions on Neural Networks”, 12-6:1411–1432, 2021.

[9] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, “A recurrent latent variable model for sequential data,” in Advances in Neural Information Processing Systems 28 (NIPS 2016), 2015.

[10] Z. Liubun, V. Nesterenko, V. Mandziy, O. Karpin, "Implementation of digital filters using neural networks when the object movement dynamics are not known," Electronics and information technologies, Is. 19, pp. 48–57, 2022.

Automating Web Scraping of User Comments for Sentiment Analysis in Social Networks

Iryna Mysiuk
Department of System Design
Ivan Franko National University of Lviv
Lviv, Ukraine
https://orcid.org/0000-0002-3641-4518

Roman Shuvar
Department of System Design
Ivan Franko National University of Lviv
Lviv, Ukraine
https://orcid.org/0000-0001-6768-4695

***Abstract* — Social networks can show the general mood of the population or groups interested in a specific topic. Machine learning will recognize people's moods in posts and automatically classify them according to preferences. Often, the problem is a lack of training data, which can now be solved using ChatGPT. Before use, the training data is preprocessed and verticalized. Web scraping from real social network data can be used as test data to evaluate neural network performance. A decision tree machine learning algorithm is used for data classification. The study's results may be of interest to analysts and page managers on social networks Facebook and Instagram.**

Keywords* — *web scraping, social networks, data processing, machine learning, data analytics

I. Introduction

According to the latest studies [1–3], people spend a large part of the day on social networks in order to quickly exchange information or share impressions while browsing the news. In general, analysts can understand news sentiments not only by the number of likes or comments, but also by the context of the comments. The text of comments will more accurately reflect the mood of users under posts on social networks. Facebook has different types of icons with emotions for posts to show the general picture about the emotions of users. In Instagram, users can also like posts, but it is impossible to understand the reaction of other users without comments. This is due to the fact that users usually do not post anything when they are not interested, given the lack of options for a quick response to posts. The described study can be a result for the analysis of information trends [4, 5].

A classification based on machine learning methods can provide quick classification of the reaction of the text in the comments. During the training phase of the model, collecting a large amount of data and labeling it as positive or negative comments is not an easy task. With a large ChatGPT product created based on the Large Language Model, a training dataset can be generated that automates data collection.

II. Methods and Tools

A. Web scraping process

The web scraping process consists of collecting text from attributes using the search for unique element locators in the Document Object Model (DOM) tree. Among the possible descriptions of locators are XPath, CSS, search by id, class or tag. Among them, XPath and CSS are considered the most reliable, considering the peculiarities of modern site development with variable class names and identifiers. In addition, frequent use of JavaScript code is required for page navigation and other actions with dynamic elements.

Access to web page elements can be automated using the Selenium library. In this way, the web driver of a particular browser can be used to access elements of a web page.

B. Used tools for text classification

The main programming language used in this work is Python, which is quite convenient for working with data due to libraries for data analysis, machine learning and visualization of results.

At the first stage, integration with the OpenAI library was used and a connection was made using a key to generate data for training. This library allows you to work both with ChatGPT in request-response mode.

Matplotlib library was used for visualization in the form of dependency graphs, pandas for working with CSV files, and numpy for mathematical operations. All processes related to natural language processing were implemented using the sklearn library.

III. Implementation

In general, the process of classifying text from comments in a social network post can be divided into the stage of preparing text data, training a neural network based on this data, and testing based on real data as shown in Fig. 1.

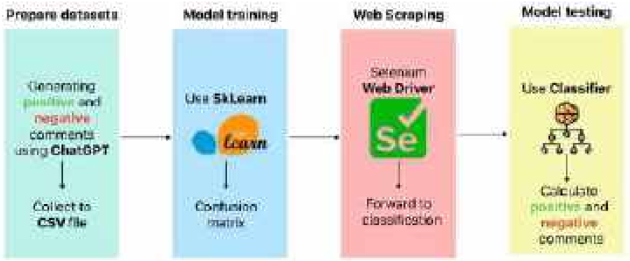


Fig. 1. Visualization of process learning and classification text

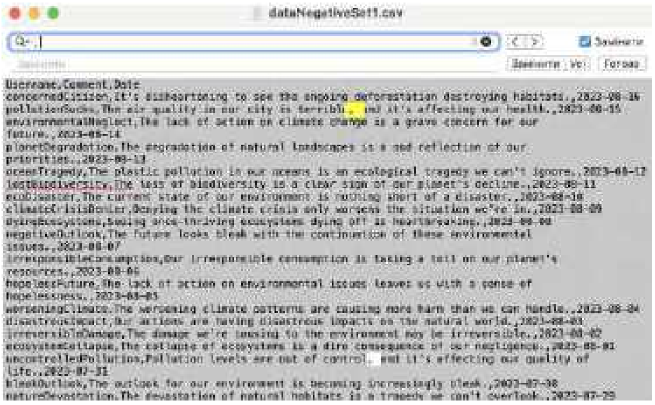
Generating data from positive and negative comments is important for the quality of learning, so it should be specific to a certain topic. In this work, the topics of environmental protection and the study of public attitudes related to their change were chosen for training and testing.

The development stage of such a system includes the training of a neural network based on the generated data. Labeled data in the file according to the principle of positive and negative comments.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

A. Generating data and preprocessing text

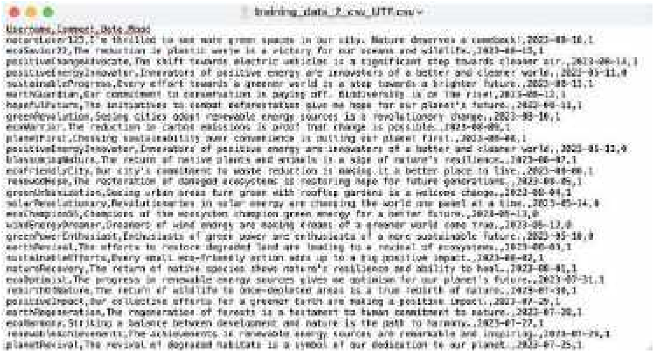
For a request in OpenAI, you need to specify engine, prompt and maximum token as parameters in the create method in Completion. Two requests were made to generate negative feedback about environmental issues and positive feedback about green energy. We collect the results in a csv file, but as shown in Fig. 2, the data must be cleaned of unnecessary punctuation marks and symbols. These characters affect the reading of the document.



```
Username,Comment,Date
concernedCitizen,It's disheartening to see the ongoing deforestation destroying habitats.,2023-08-16
pollutionSucks,The air quality in our city is terrible, and it's affecting our health.,2023-08-15
environmentalNeglect,The lack of action on climate change is a grave concern for our
future.,2023-08-14
planetDegradation,The degradation of natural landscapes is a sad reflection of our
priorities.,2023-08-13
oceanTragedy,The plastic pollution in our oceans is an ecological tragedy we can't ignore.,2023-08-12
lostBiodiversity,The loss of biodiversity is a clear sign of our planet's decline.,2023-08-11
ecoDisaster,The current state of our environment is nothing short of a disaster.,2023-08-10
climateCrisisDenier,Denying the climate crisis only worsens the situation we're in.,2023-08-09
dyingEcosystems,Seeing once-thriving ecosystems dying off is heartbreaking.,2023-08-08
negativeOutlook,The future looks bleak with the continuation of these environmental
issues.,2023-08-07
irresponsibleConsumption,Our irresponsible consumption is taking a toll on our planet's
resources.,2023-08-06
hopelessFuture,The lack of action on environmental issues leaves us with a sense of
hopelessness.,2023-08-05
worseningClimate,The worsening climate patterns are causing more harm than we can handle.,2023-08-04
disastrousImpact,Our actions are having disastrous impacts on the natural world.,2023-08-03
irreversibleDamage,The damage we're causing to the environment may be irreversible.,2023-08-02
ecosystemCollapse,The collapse of ecosystems is a dire consequence of our negligence.,2023-08-01
uncontrolledPollution,Pollution levels are out of control, and it's affecting our quality of
life.,2023-07-31
bleakOutlook,The outlook for our environment is becoming increasingly bleak.,2023-07-30
natureDevastation,The devastation of natural habitats is a tragedy we can't overlook.,2023-07-29
```

Fig. 2. Negative dataset

As in Fig. 3 shows an example of the generated data except for the comment column, the generated usernames and dates are added. For training, randomly added labels with mood as 1 positive comment and 0 negative.



```
Username,Comment,Date,Mood
natureLover123,I'm thrilled to see more green spaces in our city. Nature deserves a comeback!,2023-08-16,1
ecoSavior22,The reduction in plastic waste is a victory for our oceans and wildlife.,2023-08-15,1
positiveChangeAdvocate,The shift towards electric vehicles is a significant step towards cleaner air.,2023-08-14,1
positiveEnergyInnovator,Innovators of positive energy are innovators of a better and cleaner world.,2023-05-11,0
sustainableProgress,Every effort towards a greener world is a step towards a brighter future.,2023-08-13,1
earthGuardian,Our commitment to conservation is paying off. Biodiversity is on the rise!,2023-08-12,1
hopefulFuture,The initiatives to combat deforestation give me hope for our planet's future.,2023-08-11,1
greenRevolution,Seeing cities adopt renewable energy sources is a revolutionary change.,2023-08-10,1
ecoWarrior,The reduction in carbon emissions is proof that change is possible.,2023-08-09,1
planetFirst,Choosing sustainability over convenience is putting our planet first.,2023-08-08,1
positiveEnergyInnovator,Innovators of positive energy are innovators of a better and cleaner world.,2023-05-11,0
blossomingNature,The return of native plants and animals is a sign of nature's resilience.,2023-08-07,1
ecoFriendlyCity,Our city's commitment to waste reduction is making it a better place to live.,2023-08-06,1
renewedHope,The restoration of damaged ecosystems is restoring hope for future generations.,2023-08-05,1
greenUrbanization,Seeing urban areas turn green with rooftop gardens is a welcome change.,2023-08-04,1
solarRevolutionary,Revolutionaries in solar energy are changing the world one panel at a time.,2023-05-14,0
ecoChampion88,Champions of the ecosystem champion green energy for a better future.,2023-05-13,0
windEnergyDreamer,Dreamers of wind energy are making dreams of a greener world come true.,2023-05-12,0
greenPowerEnthusiast,Enthusiasts of green power are enthusiasts of a more sustainable future.,2023-05-10,0
earthRevival,The efforts to restore degraded land are leading to a revival of ecosystems.,2023-08-03,1
sustainableEfforts,Every small eco-friendly action adds up to a big positive impact.,2023-08-02,1
natureRecovery,The return of native species shows nature's resilience and ability to heal.,2023-08-01,1
ecoOptimist,The progress in renewable energy sources gives me optimism for our planet's future.,2023-07-31,1
rebirthOfNature,The return of wildlife to once-depleted areas is a true rebirth of nature.,2023-07-30,1
positiveImpact,Our collective efforts for a greener Earth are making a positive impact.,2023-07-29,1
earthRegeneration,The regeneration of forests is a testament to human commitment to nature.,2023-07-28,1
ecoHarmony,Striking a balance between development and nature is the path to harmony.,2023-07-27,1
renewableAchievements,The achievements in renewable energy sources are remarkable and inspiring.,2023-07-26,1
planetRevival,The revival of degraded habitats is a symbol of our dedication to our planet.,2023-07-25,1
```

Fig. 3. Full dataset

Before training, the text was verticalized using the ready-to-use TfidfVectorizer from the sklearn library. As a result, such a translation into a numerical representation of the text will help to distinguish between different texts.

B. Training model

The learning process consists in choosing a classifier (Random Forest, Decision Tree, Caussian Naïve Bayes, K Neighbors, Support Vector) and calculating the efficiency of the algorithm based on metrics. In Fig. 4 shows the differences in training based on the accuracy score metric for positive and negative feedback. As a result, the highest scores in the Decision Tree classifier are 92% different based on positive comments and 100% based on negative comments. Other classifiers have lower indicators, this may be due to implementation features.

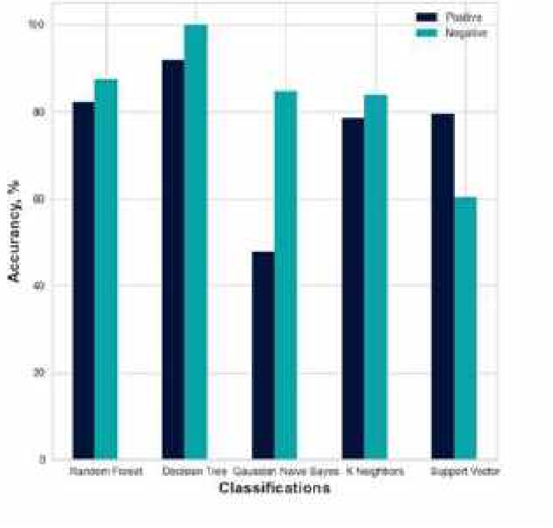


Fig. 4. Results of classification using Random Forest, Decision Tree, Caussian Naïve Bayes, K Neighbors, Support Vector

The amount of data used for training is 500 records of positive comments and 500 records of negative comments.

C. Data collection

The process of reading data consists in searching for posts on a certain topic and reading text from comments by scrolling through them. The reading process is automated using the Selenium library and tested on Instagram and Facebook social networks.

As shown in Fig. 5, a post under the tag #greenenergy is opened, the comment button is clicked (the button called "View all comments") and the texts from the comment elements are read.

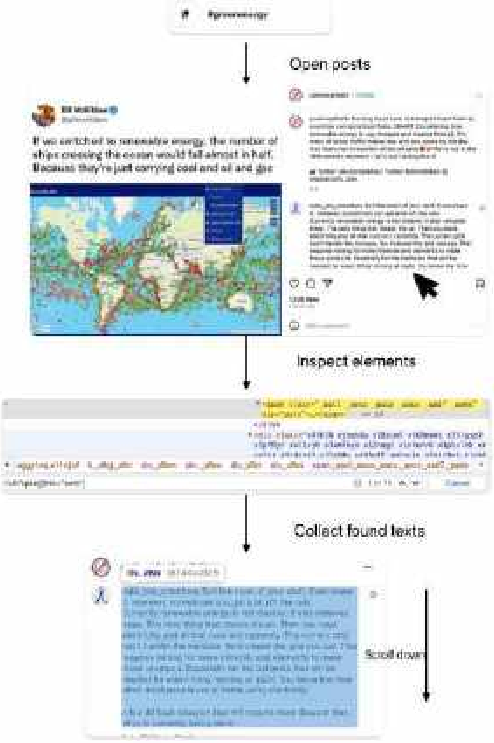

#greenenergy
Open posts
If we switched to renewable energy, the number of ships crossing the ocean would fall almost in half. Because they're just carrying coal and oil and gas
Inspect elements
Collect found texts
Scroll down

Fig. 5. Visualization of collecting process from Instagram

A similar algorithm is used to collect comments from the Facebook page as it is shown in Fig. 6. The difference lies in the different paths to the elements and buttons that are specified for assembly.

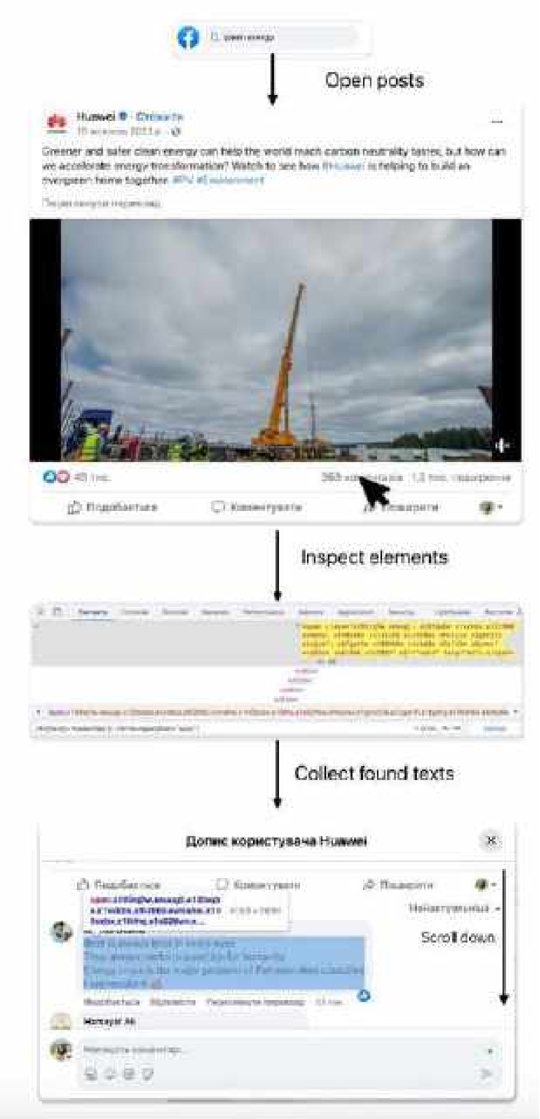


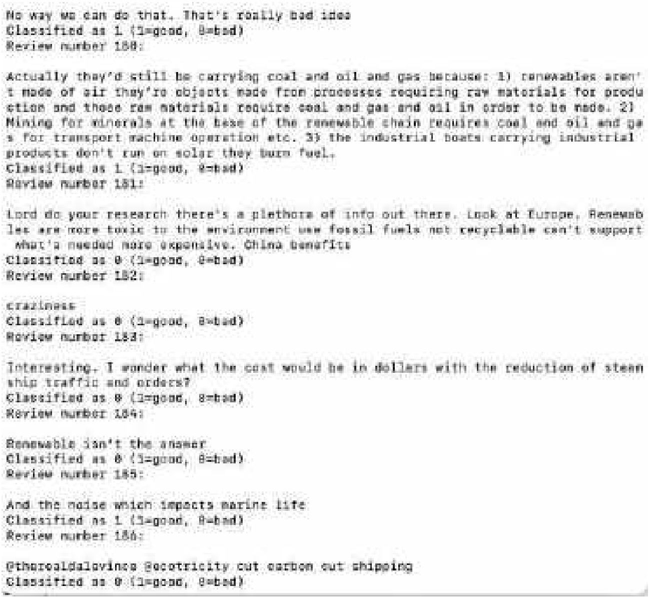
Fig. 6. Visualization of collecting process from Facebook

In this case, the data is written to a file and submitted to a neural network for sentiment recognition. However, it is possible to recognize such phrases in real time. Each time a new comment is submitted to the input of the neural network.

IV. Results and Analysis

Conducting analytics on user sentiment under the topic can be done after classifying texts based on comments under the post. As shown in Fig. 7, part of the comments from the post from the social network Instragram is classified and the information is displayed in the console view. In the same way it is done for comment of post in Facebook. All information is summed and after the last available comment will be displayed in the form of a bar chart with two columns. This dependence graph will show the ratio of the number of comments in the reference number of all posts to negative or positive.

It is best to compare the results of opposite subjects to check the adequacy of the classification. Therefore, posts under tags about environmental problems and changes related to green energy were chosen for testing. In addition to user classification analytics, you can compare the number of likes and their type to the percentage of comments classification.



```
No way we can do that. That's really bad idea
Classified as 1 (1=good, 0=bad)
Review number 180:

Actually they'd still be carrying coal and oil and gas because: 1) renewables aren'
t made of air they're objects made from processes requiring raw materials for produ
ction and those raw materials require coal and gas and oil in order to be made. 2)
Mining for minerals at the base of the renewable chain requires coal and oil and ga
s for transport machine operation etc. 3) the industrial boats carrying industrial
products don't run on solar they burn fuel.
Classified as 1 (1=good, 0=bad)
Review number 181:

Lord do your research there's a plethora of info out there. Look at Europe. Renewab
les are more toxic to the environment use fossil fuels not recyclable can't support
 what's needed more expensive. China benefits
Classified as 0 (1=good, 0=bad)
Review number 182:

craziness
Classified as 0 (1=good, 0=bad)
Review number 183:

Interesting. I wonder what the cost would be in dollars with the reduction of steam
ship traffic and orders?
Classified as 0 (1=good, 0=bad)
Review number 184:

Renewable isn't the answer
Classified as 0 (1=good, 0=bad)
Review number 185:

And the noise which impacts marine life
Classified as 1 (1=good, 0=bad)
Review number 186:

@therealdalovinco @ecotricity cut carbon cut shipping
Classified as 0 (1=good, 0=bad)
```

Fig. 7. Visualization of process classification text based on comment

These parameters may be related to each other, taking into account the fact that the number of likes in a post may be less with negative comments, and more with positive comments. The indicated trend with negative comments may be due to the lack of ability to respond to information with likes in such quick methods due to the absence of a choice of emotions

A. *Testing based on comment in post about enviromental issue*

In recent years, the process of classifying texts on the topic of environmental impacts has had a particularly negative impact on society. For example, such eco-disasters that have occurred in the world, whether it is a depressurization of a gas pipeline in the ocean, a hurricane, or a dam explosion.

If we take for analysis the information from the comments of the Instagram post shown in Fig. 5, the number of negative comments is quite large and equal to 71 percent.

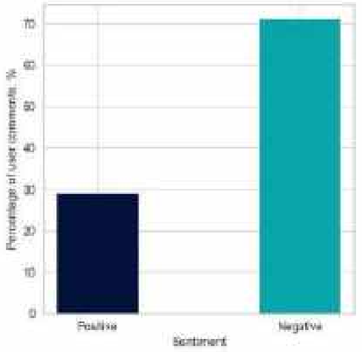


Fig. 8. Result of classification for environmental issue related to transfer of oil and gas in ocean from Instagram post

In this post, the information is more general and for most of the users who left a comment, it is outrageous. A part of the text identified as positive may contain large comments with ambiguous texts during recognition, as was the case in Fig. 7.

More interesting cases for analysis are recent events that, in addition to environmental consequences, have a global political context. To analyze the comments, a post with a video showing the consequences of blowing up the Kakhovska dam was chosen, as shown in Fig. 8.

facebook
Video
BBC News
uilding floats along flooded river after dam destroyed in Ukraine

Fig. 9. Comments from Facebook post about Kakhovska dam

After the classification of 372 comments, the result of the comments showed a lot of negative comments of 52 percent.

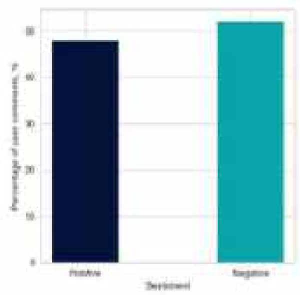


Fig. 10. Result of classification for environmental issue related to Kakhovska dam from Facebook post

However, many texts are dismissed as positive, which may be due to a lack of training data or the presence of bots.

B. Testing based on comment in post about green energy

From the Facebook post shown in Fig. 6 comments were collected and classified into two groups, positive and negative. Here is a post about green energy. 369 comments were processed in this post. The result of such classification with data is shown in Fig. 10.

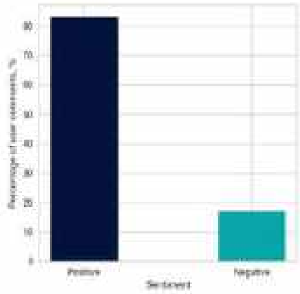


Fig. 11. Result of classification for green energy topic from Facebook post

The total number of positive comments about changes in nature with the introduction of green energy. A small part of 17 percent of users either left ambiguous comments or used negative words in the text of comments.

A post from Instagram about news about environmental things is shown in Fig. 12. In general, the news is described in a good way and comments are expected to be mostly positive.

Fig. 12. Comments from Instagram post about Kakhovska dam

The classification results are shown in Fig. 13, where 68 percent of comments are positive. This shows that this method of detecting general sentiments in the news can be used. In the future, the work should increase the data sets, perhaps this will improve the accuracy of the determination in the texts.

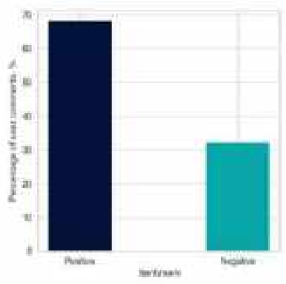


Fig. 13. Result of classification related to Good Eco News from Instagram post

C. Comparing number of likes to comment sentiments

The relationship between the number of comments and the number of likes can be, which show some things about the behavior of a person in social networks. The number of comments and likes in the social network under different posts may differ depending on the topic of the post and features of Instagram and Facebook as shown in Table 1.

TABLE I. COMPARATIVE TABLE OF POSITIVE AND NEGATIVE COMMENTS AND THE NUMBER OF LIKES FOR SPECIFIC TOPICS

Social Network	*Topic*	*Number of likes*	**Comments, %**	
			Positive	*Negative*
Instagram	Transfer of oil and gas in ocean	7026	29	71
	Good Eco News	312	68	32
Facebook	Kakhovska dam	372	48	52
	Green energy	369	83	17

For analytics, the number of likes on the page and comments can serve as an alternative. However, this information is general and it is difficult to understand why this topic is impressive. But when working with texts, you can highlight the main most frequent words from the comments of a social network post. This method can help to understand the subject or the reason for such sentiments. Facebook has different types of like icons (sad, angry and happy) that can simultaneously signal outrageous information or vice versa.

Overall, the classification can be considered successful for several examples. The disadvantage is the limited use of the trained model for other texts with different topics where the recognition is different. As an alternative to data collection for training, artificial texts from ChatGPT can be used.

The support system for making innovative business decisions regarding the implementation of investment projects [6, 7] in this direction is also important here, especially in conditions of instability, complexity and ambiguity, uncertainty and risk [8–10]. At the same time, it is necessary to take into account the peculiarities, specifics and individual aspects of the use of information systems and technologies, taking into account the methods and tools, management standards, information processing and parameters assessment criteria [11–15], existing conditions [16, 17], values of sustainable development and modeling aspects [18–22].

CONCLUSIONS

Sentiment analysis is performed based on the analysis of classified comments on social networks Instagram and Facebook posts. The training process is performed based on the generated data of positive and negative feedback from ChatGPT. The specifics of the classification are chosen for green energy as positive feedback and negative as environmental factors and will not be suitable for others. Selected real reviews from posts from similar posts by topic and automatically collected data from comments. All comments are divided into two groups, positive and negative, and dependence is shown for selected topics. Such a comparison of parameters makes it possible to analyze the general mood of users in social networks.

REFERENCES

[1] H. Cui, S. Shao, S. Niu, C. Shi, and L. Zhou, “A classification method for social information of sellers on social network,” EURASIP Journal on Image and Video Processing, vol. 2021, no. 1, Jan. 2021, https://doi.org/10.1186/s13640-020-00545-z

[2] A. Bhardwaj, “Sentiment Analysis and Text Classification for Social Media Contents Using Machine Learning Techniques,” SSRN Electronic Journal, 2020, https://doi.org/10.2139/ssrn.3735851

[3] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, “A review on sentiment analysis from social media platforms,” Expert Systems with Applications, vol. 223, p. 119862, Aug. 2023, https://doi.org/10.1016/j.eswa.2023.119862

[4] Pavlyshenko, B.M., Methods of Informational Trends Analytics and Fake News Detection on Twitter (arXiv:2204.04891). 2022. URL: https://arxiv.org/pdf/2204.04891.pdf

[5] Pavlyshenko, B.M., “Forming Predictive Features of Tweets for Decision-Making Support” (arXiv:2201.02049). 2022. URL: https://arxiv.org/pdf/2201.02049.pdf

[6] R. Skrynkovskyi, “Investment attractiveness evaluation technique for machine-building enterprises”, Actual Problems of Economics, no. 7(85), pp. 228–240, 2008.

[7] I. Mysiuk, “Designing a Data Warehouse for Collected Data About User Activity in Social Networks Using Elasticsearch,” Path of Science, vol. 9, no. 7, pp. 4001–4005, Jul. 2023, https://doi.org/10.22178/pos.94-13

[8] N. Pavlenchyk et al., “The influence of management creativity on the optimality of management decisions over time: An innovative aspect,” Journal of Eastern European and Central Asian Research (JEECAR), vol. 10, no. 3, pp. 498–514, Jun. 2023, https://doi.org/10.15549/jeecar.v10i3.1318

[9] N. Popova, A. Kataiev, A. Nevertii, O. Kryvoruchko, and R. Skrynkovskyi, “Marketing Aspects of Innovative Development of Business Organizations in the Sphere of Production, Trade, Transport, and Logistics in VUCA Conditions,” Studies of Applied Economics, vol. 38, no. 4, Feb. 2021, https://doi.org/10.25115/eea.v38i4.3962

[10] N. Popova et al., “Development of trust marketing in the digital society,” Economic Annals-XXI, vol. 176, no. 3–4, pp. 13–25, Aug. 2019, https://doi.org/10.21003/ea.v176-02

[11] R. Mysiuk, I. Mysiuk, G. Pawlowski, V. Yuzevych, M. Yasinskyi, and Y. Tyrkalo, “Video-based Concrete Road Damage Assessment Using JetRacer Kit,” 2023 17th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), Feb. 2023, https://doi.org/10.1109/cadsm58174.2023.10076528

[12] M. Babych et al., “Substantiation of economic efficiency of using a solar dryer under conditions of personal peasant farms,” Eastern-European Journal of Enterprise Technologies, vol. 6, no. 8 (84), pp. 41–47, Dec. 2016, https://doi.org/10.15587/1729-4061.2016.83756

[13] R. V. Mysiuk et al., “Determination of conditions for loss of bearing capacity of underground ammonia pipelines based on the monitoring data and flexible search algorithms,” Archives of Materials Science and Engineering, vol. 115, no. 1, pp. 13–20, May 2022, https://doi.org/10.5604/01.3001.0016.0671

[14] V. Yuzevych, O. Klyuvak, and R. Skrynkovskyy, “Diagnostics of the system of interaction between the government and business in terms of public e-procurement,” Economic Annals-XXI, vol. 160, no. 7–8, pp. 39–44, Oct. 2016, https://doi.org/10.21003/ea.v160-08

[15] L. Yuzevych, R. Skrynkovskyy, and B. Koman, “Development of information support of quality management of underground pipelines”, EUREKA: Physics and Engineering, vol. 4, pp. 49–60, Jul. 2017, https://doi.org/10.21303/2461-4262.2017.00392

[16] R. Dzhala et al., “Simulation of Corrosion Fracture of Nano-Concrete at the Interface with Reinforcement Taking into Account Temperature Change”, 4th International Workshop on Modern Machine Learning Technologies and Data Science, MoMLeT&DS 2022, CEUR Workshop Proceedings 3312, Leiden–Lviv, The Netherlands–Ukraine, pp. 123–133, Nov., 25–26, 2022, URL: https://ceur-ws.org/Vol3312/paper10.pdf

[17] A. Sumets et al., “Methodological toolkit for assessing the level of stability of agricultural enterprises,” Agricultural and Resource Economics: International Scientific E-Journal, vol. 8, no. 1, pp. 235–255, Mar. 2022, https://doi.org/10.51599/are.2022.08.01.12

[18] R. Mysiuk, V. Yuzevych, B. Koman, and M. Yasinskyi, “High Availability System for Monitoring Material Degradation Processes at the Concrete-polymer Interface,” 2022 12th International Conference on Advanced Computer Information Technologies (ACIT), Sep. 2022, https://doi.org/10.1109/acit54803.2022.9913086

[19] R. Skrynkovskyy, N. Pavlenchyk, S. Tsyuh, I. Zanevskyy, and A. Pavlenchyk, “ Economic-mathematical model of enterprise profit maximization in the system of sustainable development values,” Agricultural and Resource Economics: International Scientific E-Journal, vol. 8, no. 4, pp. 188–214, Dec. 2022, https://doi.org/10.51599/are.2022.08.04.09

[20] B. Batrinca and P. C. Treleaven, “Social media analytics: a survey of techniques, tools and platforms,” AI & SOCIETY, vol. 30, no. 1, pp. 89–116, Jul. 2014, https://doi.org/10.1007/s00146-014-0549-4

[21] K. Sailunaz and R. Alhajj, “Emotion and sentiment analysis from Twitter text,” Journal of Computational Science, vol. 36, p. 101003, Sep. 2019, https://doi.org/10.1016/j.jocs.2019.05.009

[22] N. K. Singh, D. S. Tomar, and A. K. Sangaiah, “Sentiment analysis: a review and comparative analysis over social media,” Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 1, pp. 97–117, May 2018, https://doi.org/10.1007/s12652-018-0862-8

Crowdfunding in Ukraine: Problems and Prospects for the Development of an Alternative Way of Financing Science

Nadiia Petrenko
Department of Problems of Activities and Development Strategies of the National Academy of Sciences of Ukraine
Dobrov Institute for Scientific and Technological Potential and Science History Studies of the NAS of Ukraine
Kyiv, Ukraine
ID ORCID: 0000-0002-9781-5622

Olena Vovchenko
Centre for innovations and technological development
Dobrov Institute for Scientific and Technological Potential and Science History Studies of the NAS of Ukraine
Kyiv, Ukraine
ID ORCID: 0000-0001-7502-5702

***Abstract* — This study reveals the essence of the concept of crowdfunding as an alternative way of financing creative ideas, startups, innovations, the latest technologies, and socially significant projects in the context of the development of computerization and informatization. The relationship and the difference between crowdfunding and crowdsourcing are defined. The subjects of this financing instrument are identified, and the essence of the online platforms, which are used to attract monetary allocations for the financing of the projects presented on them, is determined. The advantages for individual investors regarding investing money through crowdfunding are summarized. Domestic crowdfunding platforms have been considered and analyzed. The factors hindering the development of an alternative method of financing in Ukraine are outlined, as well as recommendations are provided for its further functioning in our country.**

Keywords* — *crowdfunding, crowdfunding platforms, innovation projects, start-ups, investors, sponsors, education, science.

I. Introduction

Currently, for the scientific community and entrepreneurs who seek to develop and use additional methods of financing for the development of their activities, to further develop their business and innovative activities. Therefore, there is an increasing need for significant financial support. This applies to both small and medium-sized enterprises and the scientific community, precisely in the situation of limited access to loans and additional financing of a specific field of activity. But in the conditions of a full-scale war in Ukraine, it has become more difficult for the scientific community, enterprises to exist and develop even with a good financial situation, it is difficult to get access to affordable loan prices and with the current state of funding of scientific institutions. In developed countries, the solution to this problem can be found in many ways through online platforms, whose influence is felt all over the world.

A long time ago, J. A. Schumpeter (1950) spoke in a similar context about "storms of creative destruction" that can take hold and begin to destroy established ways of doing science and business. Today, in developing countries, crowdfunding is one such way of financing that can change traditional business management and support the financial situation in scientific structures. In the conditions of the development of the market economy of Ukraine, the issue of finding various sources of funding is becoming more and more important for both scientists and entrepreneurs [1].

According to H. Le Bon (1895). While all our ancient beliefs are shaking and fade away, while the old pillars of society are given way one after another, the power of the mob is the only power which is not threatened, and whose prestige is ever increasing. The age we are about to enter will indeed be the age of the crowd" [2].].

II. Methods of investigations

For the scientific justification of the results of crowdfunding research and crowdfunding platforms as alternative financing in Ukraine, a statistical, theoretical and generalization method was used. Such general scientific methods and techniques as scientific abstraction, grouping, classification, comparison, induction, deduction, analysis, synthesis and others are applied directly in the research process.

III. Results and Discussion

Due to the current unstable situation in Ukraine, it is necessary to use progressive ways of development in order to improve the economy. For the creation and development of new business ideas, projects, startups, first of all, financial resources are needed. Therefore, there is a need for the newest methods of investing, one of them is crowdfunding. Crowdfunding is an alternative form of financing, a powerful tool for attracting financial resources.

Today, the collective effort of individuals pooling their resources, usually via the Internet, to support efforts initiated by other people or organizations is defined as crowdfunding (De Buysere et al., 2012). Crowdfunding is a disruptive financial intermediation technology with a number of challenges that politicians cannot afford to ignore. Many authors pay attention to the nature of crowdfunding and explore the main conditions of its use in different ways (Table 1) [3].

In general, we define crowdfunding as an opportunity to attract financial funds, resources of the "crowd" via the Internet. The subject of this study is the study of the main conditions for the development of crowdfunding throughout the world, the needs and prospects of its use in Ukraine.

Crowdfunding is the collective effort of a large number of people pooling small amounts of capital to fund a new or existing business or scientific community. For each campaign, the institution sets a target amount of money and a fixed period of time, each day is counted, and the collected

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

money is counted so that visitors can follow its success (Statistics Portal).

TABLE I. Theoretical approaches regarding the essence of crowdfunding

№	Author	Definition of crowdfunding
1	S. Moeller [4]	Customers involved in crowdfunding are not only integrated at the stage of service provision, but also contribute to the development and customization of the offer.
2	A. Ordanini [5]	Requesting financial resources online and offline in exchange for a reward offered by the creator, such as recognition, experience, or product. This is an initiative aimed at attracting funds for a new project by collecting small and medium-sized investments from other people (crowd).
3	A. Schwienbacher, B. Larralde [6]	An open solicitation, preferably via the Internet, for financial resources in the form of donations or in exchange for some other form of reward and/or voting rights to support an initiative for specific purposes. Crowdfunding can be seen as a combination of the concepts of crowdsourcing (i.e. seeking funding from the crowd) and microfinancing (small contributions; no collateral).
4	P. Belleflamme, T. Lambert, A. Schwienbacher [7]	An open appeal over the Internet for financial resources in the form of a monetary donation, sometimes in exchange for a future product, service, or reward.
5	M. Poetz, M. Schreier [8]	Crowdfunding draws inspiration from concepts such as microfinance. Crowdfunding represents its own unique category of fundraising, facilitated by a growing number of Internet sites dedicated to the topic.
6	Z.J. Griffin [9]	Crowdfunding essentially involves a sequence of processes by which a scientist or entrepreneur publishes a request for funding on a crowdfunding platform or website with a description of the proposed project.
7	G. Burtch, A. Ghose, S. Wattal [10]	All crowdfunding operations are done through online crowdfunding platforms, which also provide convenient facilities for all fund exchanges.
8	C.S. Bradford [11]	Crowdfunding can be presented differently according to the current situation.
9	O.M. Lehnera [12]	In the context of social entrepreneurship, crowdfunding is praised in the media for its multifaceted potential.
10	R. Wash [13]	Crowdfunding can be done in many ways - through an open call on a web page, by posting an announcement in a public place, or through an organized online marketplace called a crowdfunding website.

For Ukraine, crowdfunding can be more effective than traditional methods of financing, but for this it is necessary to learn from already developed world experience to understand how crowdfunding works.

The development of crowdfunding in Ukraine is possible with specific strategies. According to World Bank research, they can be defined as economic, social, technological and cultural strategies. Economic strategies are represented by crafting exemptions from securities rules that allow for easy registration of shares, strategically linking crowdfunding with patriotic and cultural ideas. Social strategies involve top social media, experts/bloggers, "other tastemakers" who engage with audiences; media and educational activities that promote awareness; regular crowdfunding events with trusted third parties to teach successful practices. Technology strategy refers to lessons learned from the developed world to identify gaps in existing technologies for online financial transactions. Cultural strategies are represented by existing incubators/accelerators/other co-working spaces as centers for funding innovation, promoting the trust of professional investors and consumers in crowdfunding through open communication.

Every year, the results grow thanks to crowdfunding activities. In order to spread this way of investing, it is important to have easy access and understanding of aspects of crowdfunding activities. This will give an opportunity to gain trust from potential investors. Therefore, in our time, this topic is more than relevant for consideration.

Crowdfunding platforms are guarantors of the integrity of newly created projects and organizations, but there are still not a large enough number of them operating today. Crowdfunding platforms can be used not only to attract external funding, but also to test a business idea for viability and demand in the economic market, to enter foreign markets, and also as an informational opportunity to attract the attention of the media, sponsors, investors and business incubators to projects Crowdfunding, even if it is institutional, should be used as an additional source of income for the organization or external capital to start the project [14].

Since the beginning of the full-scale war in Ukraine, the use of crowdfunding platforms has become more and more relevant. First of all, the purpose of creating projects is charitable contributions from users who want to support ideas aimed at overcoming the consequences of military actions in Ukraine. Since on these platforms, funds are being collected to finance such projects as: "FluRArium - leaflet for refugees", "Etis.help - delivery of humanitarian aid", "School of aesthetic education of children of forcibly displaced people" on Spilnokosht. On the platform "Dobro.ua - Ukraine above all!" This project was created for the purpose of collecting charitable contributions from users who wish to support projects aimed at overcoming the consequences of military operations in Ukraine, including helping military personnel and civilians affected by military operations. On the "My City" platform, all collected funds are directed to the defense of the city of Odesa. The total amount of collected funds currently amounts to more than 89 thousand UAH. The funds are directed to the purchase of medicines, food products for the Armed Forces, hospitals and maternity homes (Table 2).

TABLE II. Ukrainian crowdfunding platforms

Name	Description of the platform
Spilnokosht [15]	One of the most successful crowdfunding platforms in Ukraine is «Spilnokosht», which was created in 2012 by the online site biggggidea.com. It was created in 2009 with the aim of exchanging ideas between enterprising people, that's why it's called «Big Idea». The platform hosts projects in the fields of health care, education, literature, sports, music, science, professional travel and journalism. With the help of this platform, people mainly finance festivals, public television, radio, documentary and medical projects, urban innovations.
Na-Starte [24]	This crowd platform was created by an Odesa IT company. Started working on February 1, 2014. Its main goal is the development of cultural ideas and innovations in Ukraine. Today, there are more than a hundred projects that have been launched on this platform, 15% of which have found the necessary funds, having collected more than 100,000 UAH. Despite the fact that it is an Odesa platform, projects from Zaporizhzhia,

	Kyiv, Cherkasy, Dnipro, etc. are also launched on it. Using the platform Na-Starte 3.7 million UAH (124% of the declared collection amount) were collected for the filming of Georgy Deliyev's film "Odessa Znaida".
GoFundEd [18]	Created by the public organization Center for Innovative Education "Pro. Svit". The platform was created to implement educational ideas. The authors of the projects are mostly teachers, students, and public activists. GoFundEd since 2016, has been providing school teams with mentoring support and support in the implementation of projects in schools, teaches cooperation, communication and transparent fundraising. Since the existence of the platform, more than 200 projects have been published, of which more than half of the projects have been successfully financed and about 3.5 million UAH have been received. for their financing.
RazomGo [17]	This platform was created in 2018. With its help, you can attract financing for projects in various fields: education, sports and health, art, games, design, etc. Also, this platform offers special training, which includes checklists and step-by-step instructions for starting the project and support from a curator who will always give advice and help to complete everything clearly and efficiently.
Dobro.ua [19]	This platform was founded in 2011 by the International Charity Fund "Ukrainian Charity Exchange", the founder and initiator of which was the Viktor Pinchuk Foundation. On August 12, 2020, the foundation carried out a complete update of the largest online charity platform of Ukraine, which was named Dobro.ua About 449 million UAH were collected on the platform during its long-term work. and almost 6,000 projects were supported. On the website, you can track the progress of fundraising for social projects in various spheres of life: medicine, education, culture, ecology, as well as support for local community initiatives.
My city [20]	This is a local crowdfunding platform that was created in Odessa in August 2015. The purpose of its creation was to improve the standard of living of citizens and implement local social projects in such cities as: Kharkiv, Odesa, Dnipro. Over 3.6 million UAH were collected during its operation on the platform. and more than 100 projects were successfully financed.
StartEra [26]	StartEra – an innovative crowdfunding platform for hosting and promoting creative, social and technological projects, as well as public initiatives with the financial support of those who support and cheer for the development and implementation of innovative projects. Platform founders: Lviv Polytechnic National University, Lviv Startup School, "Your City" media hub. The platform focuses on socially significant projects and positions itself as one that helps you create your own product without using loans.
Komubook [25]	Komubook – the first Ukrainian platform that works on the principle of reward for contribution. The platform selects a book worth publishing and evaluates its value. If the project is approved, it is published on the website and a fundraiser is announced, which can last from 30 to 60 days. When the required amount is accumulated, the book is issued and sent to all bakers. If the required amount of funds is not collected, then the platform has the right either to finance the publication with its own funds and still publish the book, or to oblige to return the funds to all backers. A backer is a person or organization that financially supports a project (team) during a crowdfunding campaign.

One of the famous crowdfunding platforms is the Internet platform "Kickstarter" – this is a site for financing creative projects under the shared cost scheme. The founder is the USA, the platform won the award National Design Awards. The number of subscribers is 981,383. The stated mission of the company is to help implement creative projects. KickStarter funds a variety of projects in 13 categories: art, comic, dance, design, fashion, film and video, food, video games, music, photography, publishing, technology, theater. As of 2022 KickStarter received almost 6 billion dollars from 20 million supporters to finance 205 thousand projects [3; 22; 23]. The platform is open to supporters from anywhere in the world and to creators from many countries (Table 3, Table 4) [21].

TABLE III. Ukrainian projects on the KickStarter platform

№	Project name	The final product	Funds raised, thousands $
1.	Petcube	Gadget for remote monitoring of pets	251
2.	LaMetric	A universal clock that, in addition to the time, shows other useful information from the Internet	370
3.	iBlazr	Flash for smartphones	56
4.	FORCEemotion	Asmart bracelet that tracks physical condition	30.314
5.	Phonster	Phone holster	Under consideration (approval)
6.	KrakenFix	Ski attachment, on the shoulders	10
7.	GreenNanny	A device that provides individual watering of plants	5.717
8.	Planexta	A smart bracelet that tracks your emotional state	130
9.	GearEye	System for finding lost things	558

The platform was created to implement educational ideas. The authors of the projects are mostly teachers, students and public activists.

TABLE IV. Ukrainian campaigns that started their career on Kickstarter

№	Campaign name	Orientation (what does he do)	City/ Year of establishment/ The founders	Final products
1.	Ugears	Production of wooden mechanical 3D structures	Kyiv 2014 Gennady Shestak	Ukrainian constructor for adults and children
2.	Emotion Labs	Development of methods of interpreting signals from the human body and devices using these methods.	Kyiv 2014 Elizaveta Voronkova; Ilya Kuharenko	- EMwatch (a smart watch for corporate clients with wrist pressure measurement technology); - EMtracker a device (namely, a bracelet) for recognizing stressful conditions in loved ones.

A crowdfunding platform whose main goal is the development of cultural initiatives and innovations in Ukraine. The biggest success of the platform so far is the fundraising in early 2017 for a feature film "Odessa Znayda" by the director Georgy Deliyev (3.7 million UAH were collected out of the required 3 million UAH).

A popular practice of some startup schools is to help prepare a project for placement on a crowdfunding platform. The first school that announced itself is Tech StartUp School – startup school founded by Lviv Polytechnic.

Among the successful projects, most raise from 1,000 to 9,999 dollars. These dollar amounts fall under the following categories: Design, Gaming, and Technology [22].

Conclusions

So, crowdfunding platforms are one of the newest tools for attracting investments, which is developing and gaining popularity nowadays. This economic category serves as an alternative financing for individuals for various projects or development of business ideas. This method of investing has a fundamental feature that is not typical of financial investments - it is open to everyone.

Features of crowdfunding are a high level of accessibility, since the main platform for financing is an Internet platform, that is, there is no need for intermediaries; reduction of both financial and time costs for searching for investors or projects for financing; there are no territorial restrictions on fundraising.

The conducted analysis allows us to claim that crowdfunding in Ukraine shows a stable growth in the development of science, the popularization of scientific discoveries, and scientists. Ideas for startups are increasing, and the number of investors who are ready to finance projects of various directions is also increasing.

Having analyzed the above information, it is possible to highlight the following factors that do not provide an opportunity for the development of crowdfunding in Ukraine:

- lack of a legal framework for regulating crowdfunding;

- the total amount of funds raised by Ukrainian crowdfunding platforms is insignificant compared to international ones;

- low level of knowledge about the implementation of investment activities based on crowdfunding platforms;

- distrust of this method of financing among the population, as it is a new alternative for financing new projects;

- limited solvent demand for financing projects due to the low level of income of the population;

- on these platforms, more projects are financed in such fields as education, health care, culture, etc.

Also, after analyzing current crowdfunding trends, we can conclude that the closed nature of investing in the scientific field and business will change rapidly, as the social network affects the flow of both information and capital to scientific institutions and companies. Crowdfunding provides various benefits to a wide range of users. This is due to its flexibility, community involvement and the variety of forms of funding it can offer. The development of crowdfunding as a more distributed method of capital formation is consistent with changes in the flow and distribution of information and the creation of new production capacities. Due to its limited size, crowdfunding cannot solve all financial problems by itself. But at the same time, crowdfunding is an alternative form of financing that can complement traditional financing. The growth rate of crowdfunding in both developing and developed countries indicates that it is capable of becoming a financial instrument in most countries of the world.

The topic of crowdfunding in Ukraine is still quite new and underdeveloped, which has specific problems that hinder its development. Looking at this, we can offer the following recommendations regarding the functioning of crowdfunding in Ukraine:

- formation of legislation and the corresponding legal framework in the field of crowdfunding;

- improve methods of communication between investors and project developers;

- increase the level of trust by ensuring more transparent collection of funds and ease of their transfer;

- creation of consulting services in each city, which will inform about the procedure for financing and help in the design of projects.

References

[1] J. A. Schumpeter, Capitalism, Socialism and Democracy. 3rd ed. London: Allen and Unwin, 1950, 460 p.

[2] G. Le Bon, The Crowd: a Study of the Popular Mind. London: T. Fisher Unwin, 1895, 160 p.

[3] N. S. Petrenko, “Crowdfunding as an alternative form of financing the development of science in Ukraine”, In XV International Scientific and Practical Conference Problems, priorities and prospects of sustainable development in the 21st century, Kamyants-Podilsk, May 11, 2023, pp. 23-27. [in Ukrainian].

[4] S. Moeller, “Customer integration. A key to an implementation perspective of service provision”, Journal of Service Research, vol. 11(2). pp. 197–210.

[5] A.Ordanini, “Crowd funding: customers as investors”, The Wall Street Journal, 23 March 2009.

[6] A. Schwienbacher, B. Larralde, “Crowdfunding of small entrepreneurial ventures”, The Social Sciences Research Network, 2010. [Online]. Available: papers.ssrn.com. [Accessed: Aug. 1, 2023].

[7] P. Belleflamme, T. Lambert, A. Schwienbacher, “Crowdfunding: Tapping the Right Crowd”, In International Conference of the French Finance Association (AFFI), May 11–13. [Online]. Available: www2.dse.unibo.it. [Accessed: Aug. 1, 2023].

[8] M. Poetz, M. Schreier, “The value of crowdsourcing: can users really compete with professionals in generating new product ideas?”, Journal of Product Innovation Management, Vol. 29.

[9] Z. J. Griffin, “Crowdfunding: Fleecing the American Masses”, The Social Sciences Research Network, 2012. [Online]. Available: papers.ssrn.com. [Accessed: Aug. 1, 2023].

[10] G. Burtch, A. Ghose, S. Wattal, “An Empirical Examination of the Antecedents and Consequences of Investment Patterns in Crowd-funded Markets”, The Social Sciences Research Network, 2012. [Online]. Available: papers.ssrn.com. [Accessed: Aug. 1, 2023].

[11] C. S. Bradford, “Crowdfunding and the Federal Securities Laws. Columbia Business Law”, The Social Sciences Research Network, 2012. [Online]. Available: papers.ssrn.com. [Accessed: Aug. 1, 2023].

[12] O. M. Lehnera, “Crowdfunding social ventures: a model and research agenda”, Venture Capital: An International Journal of Entrepreneurial Finance, vol. 15(4). pp. 289–311, 2013.

[13] R. Wash, “The Value of Completing Crowdfunding Projects”, In Seventh International AAAI Conference on Weblogs and Social Media, 2013. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/download/14388/14237/17906. [Accessed: Aug. 1, 2023].

[14] N. S. Petrenko, “Prospects for the use of crowdfunding in Ukraine”, In 5th International Scientific and Practical Internet Conference Integration of Education, Science and Business in Modern Enviroment: Summer Debates, Dnipro, August 3-4, 2023. [in Ukrainian] (unpublished).

[15] Crowdfunding: essence, state and peculiarities of activity in Ukraine. [Online]. Available: https://www.businesslaw.org.ua/craundfunding-v-ukraini-t/. [Accessed: Aug. 3, 2023]. [in Ukrainian].

[16] Official website “Great idea”. [Online]. Available: https://biggggidea.com. [Accessed: Aug. 2, 2023]. [in Ukrainian].

[17] Official website “RazomGo”. [Online]. Available: https://razomgo.com. [Accessed: Aug. 2, 2023]. [in Ukrainian].

[18] Official website “GoFundEd”. [Online]. Available: https://gof.org.ua. [Accessed: Aug. 2, 2023]. [in Ukrainian].

[19] Official website “dobro.ua”. [Online]. Available: https://dobro.ua. [Accessed: Aug. 2, 2023]. [in Ukrainian].

[20] Official website “My City”. [Online]. Available: https://mycity.com. [Accessed: Aug. 2, 2023]. [in Ukrainian].

[21] Official website “Kickstarter”. [Online]. Available: https://www.kickstarter.com. [Accessed: March 16, 2023].

[22] N. S. Petrenko, “The online platform "Science and Business" as a factor in the popularization of science in Ukraine during the war”. In International scientific and practical conference Modern problems of science, education and society, Kyiv, March 26-28, 2023, pp. 833-838. [in Ukrainian].

[23] N. S. Petrenko, “A modern view of the popularization of science in Ukraine in the Internet space: specifics and problems of understanding”. Bulletin of the Kamianets-Podilskyi National University named after Ivan Ohienko. Economic sciences, vol. 16, 2021, pp. 130-135. [in Ukrainian].

[24] Official website “Na-Starte”. [Online]. Available: http://ww1.na-starte.com/ [Accessed: Aug. 3, 2023]. [in Ukrainian].

[25] Official website “Komubook”. [Online]. Available: https://komubook.com.ua/ [Accessed: Aug. 3, 2023]. [in Ukrainian]

[26] Official website “StartEra”. [Online]. Available: https://startera.org.ua/ [Accessed: Aug. 3, 2023]. [in Ukrainian].

Modeling of Bank Performance Indicators Based on Business Intelligence and Data Analysis

Liudmyla Koliechkina
Algorithms and Databases Department, University of Lodz, Narutowicza 68,
Lodz, Poland
0000-0002-4079-1201

Tetiana Hudz
Department of Finance and Banking Poltava University of Economics and Trade, Poltava, Ukraine
0000-0002-2310-5425

Vadym Kylnyk
Department of Human Resources Management, Labour Economics and Economic Theory
Poltava University of Economics and Trade, Poltava, Ukraine
0009-0007-0020-6470

***Abstract* — Given the high level of uncertainty in the economic situation, ensuring the stable operation of the banking industry requires the expansion of mathematical tools for modelling the development of the situation and decision-making. The purpose of our study is to investigate the factors that influence the productivity of the banking industry and its profitability based on business intelligence and data analytics. The goal is achieved by solving the following tasks: systematization of theoretical and practical principles of using business intelligence in the banking industry; development of economic and mathematical models of the relationship between efficiency, risk and liquidity of banks; modelling the state of the banking industry of Ukraine.**

Based on a comprehensive literature review and collected data, a number of parameters were analyzed and theoretical models were developed to study the impact of key factors on the efficiency of Ukrainian banks, which allowed for the expansion of the use of business intelligence and data analytics in the banking industry.

The study used correlation and regression analysis, decision-making methods, and a statistical package for data analytics. The study revealed correlations between banks' performance and a number of indicators that characterize their financial stability, liquidity, solvency, credit and currency risks.

This study shows that it is very important to successfully plan business analytics and data analytics in order to get all the benefits of this technology and its application in the current conditions of the banking industry.

Keywords* — *business intelligence, data analytics, decision making, mathematical model, information technology, banking industry

I. Introduction

Business intelligence and data analytics are two vital techniques that today's banking systems use to capture data for further analysis.

Both of these techniques help you visualize, analyze and understand data related to your business, customers, competitors and industry.

It can help a manager make better business decisions, develop a fruitful strategy, improve their operations, increase sales and revenue, find patterns, and predict future moves.

Although business intelligence and data analytics play a crucial role and can be used interchangeably in different fields, the terms mean different things to different industries. Business intelligence and analytics (BIA) is considered one of the most critical technologies, systems, practices, and applications that help organizations develop a deeper understanding of business data and gain a competitive advantage while improving operations and product development and strengthening relationships with customers [1,2].

BIA has an even more important role in the banking sector by enabling experts and managers to make better, accurate, timely, and relevant decisions so as to increase the productivity and profitability of the bank and be able to comply with the different regulatory and environmental dimensions of this sector [3].

BIA, nowadays, is a trendy issue and a compulsory prerequisite for creating an outstanding corporate image, which goes in line with implementing a successful plan regarding using technology extensively. Thus, this supports business decisions and gains a competitive advantage in today's dynamic environment, which requires outstanding efforts for dedicating massive budgets to research and development.

Data are a focal point and are considered the fuel of the future since they can be processed efficiently and used effectively in supporting risky occurrences and decisions that can be heavily reflected in the performance of corporates [4].

Business intelligence (BI) is an umbrella term that includes structures, tools, databases, applications, and methodologies to analyze data by converting raw data into meaningful and helpful information to support business managers' decisions [3, 4].

Statistical methods of database processing in combination with mathematical tools are used quite actively to improve the work of banks. There are three main areas of studying banks' activities by means of mathematical analysis. The first direction involves econometric modelling of bank performance under the influence of a certain range of factors [5, 6, 7]. Along with this, the method of correlation and regression analysis is used to forecast bank profits as an indicator of its effective operation [8]. There are examples of building models to assess the impact of external factors on bank performance [9, 10, 11]. However, today's realities require the implementation of similar studies with the reformatting of the composition of factors influencing the activities of banks in accordance with existing risks and threats.

The second area is related to the analysis of banking risks, namely operational risk [12], credit and liquidity risk [13], and market risk on the example of customer outflows from banks [14]. Particularly noteworthy are the scientific and practical results obtained through the design of neural networks for assessing the credit risk of bank borrowers [15, 16]. The

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

dynamism and innovation of the banking business creates an open need to constantly improve the quality of its cybersecurity - the third area of analysis of bank databases. This seems to be possible by building mathematical models [17, 18].

The modern problem of ensuring the efficiency of banks is multifaceted. Therefore, its in-depth study requires the latest research using mathematical analysis methods to build adequate models for assessing the main financial contradiction in the activities of banks under increased economic uncertainty.

This work presents a study of factors affecting the productivity of the banking sector and its profitability based on business analytics and data analysis, and presents a mathematical model of choice in the field of bank lending based on decision-making theory.

Such studies were conducted in a number of both foreign and domestic literary sources and are relevant today.

Based on a comprehensive literature review and collected data, a number of parameters were analyzed and theoretical models were developed to study the influence of key factors on the implementation of business analytics and data analysis in the banking sector.

The study used correlation-regression analysis, decision-making methods, and a statistical package for data analysis. As a result of the study, dependencies between a number of indicators and their impact on the efficiency of banks were revealed.

This study shows that it is very important to successfully plan for business intelligence and data analysis in order to get the full benefits of such technology and its application, especially in the banking sector.

The work is organized as follows - the following section presents a review of key related literature, and, after that, Section 2 explains the theoretical framework and hypotheses development. Data collection and methodology are presented in Section 3, whereas Section 4 presents the research results. The last section presents the conclusion along with the limitations and directions for further research.

II. Literature Review and Background

A. Business Intelligence and Analytics

The term business intelligence (BI) was popularized during the 1990s and could be considered a term encompassing a wide variety of processes and software used to collect, analyze, and disseminate data in the interest of better decision making [1, 2].

Banks have always been the leading organizations in using the latest technologies, applications, and tools that can improve their business or increase productivity, profits, sales, or give them a competitive advantage among competitors. Similar to other technologies, business intelligence and analytics promises to help the bank acquire much more and better insights than the classical report technologies and provide more accurate and precise data analyses. In fact, BIA enables them to raise both operational and management levels of data underrating and analysis that can lead to increased sales and profits [1-3].

B. Problems of decision-making in the banking sector and methods of solving them

Along with various research techniques and methods of bank analysis, business models and mathematical decision-making models are quite effective. Such models and methods are considered in detail in the work [19, 20] of others. Here, more interest in decision-making problems with many criteria, that is, problems of multicriteria optimization. Such problems have applications in various fields [20, 21].

Multicriteria problems and methods are widely used for decision-making in various commercial and financial contexts because of the diversity of solutions they can provide. In many studies in which financial decision-making problems have been evaluated, financial decisions have been shown to be multidimensional [19-22]. Hence, most scientists and practitioners apply the methods of multicriteria operations research when solving financial decision-making problems. Research in the field of multicriteria optimization is currently particularly intensively stimulated by practical needs and the development of computer information technologies.

The problem of decision-making in the economy and finance, in particular optimal planning, arises due to two fundamental circumstances: on the one hand, the multivariate nature of planning decisions, on the other hand, the purposefulness of economic and financial systems. The set of alternative plan options is determined by the available opportunities for economic development; and the selection from this set is the goal of the system to be planned. The adopted decision is the result of a joint consideration of the goals and the possibilities of their coordination with each other.

When using mathematical methods in the analysis and making of planning decisions for financial institutions, both components of the selection problem should find an adequate reflection in the economic-mathematical model.

The system of goals appears everywhere and in those situations where the assumption of the existence of a single objective function does not seem excessive. As a result, the selected partial goals can be assigned specialized financial indicators as criteria with sufficient grounds, which are reflected in applied models without much difficulty. The multiplicity of goals of financial systems has an objective nature and finds its model reflection in the form of a vector criterion, and therefore a multi-criteria problem. Therefore, let the elements of the goal system be represented by partial criteria $f_1,...,f_l : X \to R^l$; here and further X is a set of admissible plans, identified with its economic-mathematical description and located in the space R^n . The functions $f_1,...,f_l$ form a vector optimality $F(x)=(f_1(x),...,f_l(x))$ criterion; it is assumed that the j -th goal corresponds to the maximization of the component $f_j(x)$. The pair (F,X) formed by the vector criterion $F : X \to R^l$ and the set X is a model or problem of multicriteria (vector) optimization, which we will denote by $Z(F,X)$.

Next, the presented model will be defined for finding financial indicators and their further analysis.

III. Methodological Aspect

A. Sample Study

The study population consists of Ukraine commercial banks that use business intelligence and analytics on a large scale.

In the second year of the full-scale invasion, Ukrainian banks are providing services without interruption, maintaining network operations, maintaining operational efficiency, profitability and increasing capital. Banks are currently well provided with liquid funds. On average, banks' liquidity ratios are three times higher than the minimum requirements. The amount of households' deposits with banks is stable, and the share of time deposits has started to grow. Businesses continue to attract funds, while the share of refinancing loans and external borrowings in banks' liabilities has fallen to a minimum.

The business loan portfolio continues to decline due to weak demand resulting from insufficient solvency to service the loans. The retail loan portfolio stabilised after a deep decline due to higher demand for card loans to meet current needs. Sporadic surges in mortgage lending are supported by government programmes to help Ukrainians rebuild their damaged housing. However, funding for these programmes is not systematic. Uncertainty hinders the development of the real estate market and mortgage lending.

Despite the losses from the war, banks were profitable in 2022, and in 2023, profits increased. The high interest margin makes banks comfortable with a possible decline in interest rates, so the risks to profitability are being smoothed out. The banks' profitable operations have helped maintain their capital adequacy. Currently, it is twice the minimum requirement. With their accumulated profits, banks will continue to have to meet deferred and new capital requirements under European regulations.

B. Data Base

The research is based on a database of 65 Ukrainian banks as of 01 June 2023. The database covers economic standards of banks' activities, as well as their financial statements. The input data base for business analytics is based on information published on the official website of the National Bank of Ukraine [23].

IV. Economic-mathematical Modeling and data analysis

A. Model (1)

The economic-mathematical model (1) of efficiency of banks in Ukraine has the form:

$$\begin{aligned}Y_1 &= 0.32504 + 0.08211 \cdot X_1 - 0.08881 \cdot X_2 + 0.00333 \cdot X_3 + \\ &+ 0.00276 \cdot X_4 + 0.20019 \cdot X_5 - 0.00007 \cdot X_6 + 0.00020 \cdot X_7 - \\ &- 0.00068 \cdot X_8\end{aligned} \tag{1}$$

where Y_1 is the return on assets of the bank;

X_1 – bank regulatory capital adequacy ratio;

X_2 – core capital adequacy ratio;

X_3 – maximum exposure to credit risk per counterparty;

X_4 – risk ratio of the total long open currency position;

X_5 – risk ratio of the total short open currency position;

X_6 – liquidity coverage ratio for all currencies;

X_7 – foreign currency liquidity coverage ratio;

X_8 – net stable funding ratio ratio.

Figure 1 shows the actual distribution of Ukrainian banks by return on assets relative to the 1.56 threshold.

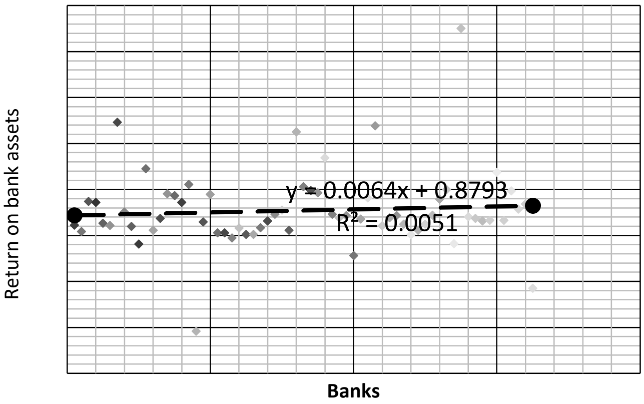


Fig. 1. Breakdown of Ukraine banks by return on assets as at 2023.01.06 (source: authors on the base of [19] data)

The marginal value (1.56) of the return on assets of banks is determined on the basis of model (1), taking into account the thresholds of the X-factors: X_1=10%, X_2=7%, X_3=25%, X_4=5%, X_5=5%, X_6=100%, X_7=100%, X_8=100%. Most banks tend to be close to the marginal return on assets. Sharp deviations from the return on assets threshold indicate that the bank is taking on significant risk.

Second Constraint (2) for Y = 0.32504. It is equal to the free term and means what Y will be if all X-factors are equal to zero.

B. Characteristics of the model (1)

According to the Fisher's criterion, it can be argued that model (1) is reliable with a probability of 0.99. This is proved by the following condition: the actual value of Fisher's criterion 4.46307 exceeds its tabulated value at the level since the actual value is 2.84694.

Parameters of the model (1) are shown in the Table 1.

TABLE I. Characteristic of The Economic-mathematical Model (1)

Factor	Weight values of the model coefficients	Standard errors of the coefficients of the variables	T-test for the model coefficients	Elasticity coefficients
X_1	0.08211	0.02337	3.51380	3.93693
X_2	-0.08881	0.02387	-3.72017	-3.39231
X_3	0.00333	0.03241	0.10288	0.04909
X_4	0.00276	0.00571	0.48334	0.01922
X_5	0.20019	0.16755	1.19484	0.14490
X_6	-0.00007	0.00006	-1.17170	-0.06996
X_7	0.00020	0.00009	2.28142	0.17760
X_8	-0.00068	0.00173	-0.39223	-0.16354

[a] calculated by the authors

Let us describe the parameters of model (1).

1. With an increase in the regulatory capital adequacy ratio (X_1) by 1, with other factors remaining constant, the return on assets of the bank (Y_1) increases by 0.08211 on average.

2. An increase in the core capital adequacy ratio (X_2) by 1, with all other factors held constant, leads to a decrease in the bank's return on assets (Y_1) by 0.08881 on average due to the creation of provisions.

3. An increase in the maximum credit risk exposure per counterparty (X_3) by 1, with other factors remaining constant, leads to an increase in the return on assets (Y_1) by an average of 0.00333. Risky transactions potentially bring higher profits.

4. An increase in the risk ratio of the total long open currency position (X_4) by 1, with other factors remaining constant, leads to an increase in the return on assets of the bank (Y_1) by an average of 0.00276.

5. An increase in the risk ratio of the total short open currency position (X_5) by 1, with other factors remaining unchanged, leads to an increase in the return on assets of the bank (Y_1) by an average of 0.20019.

6. An increase in the liquidity coverage ratio for all currencies (X_6) by 1, with other factors remaining constant, leads to a decrease in the return on assets of the bank (Y_1) by an average of 0.00007. Excessive liquidity always reduces the bank's profit.

7. An increase in the foreign currency liquidity coverage ratio (X_7) by 1, with other factors remaining constant, leads to an increase in the return on assets of the bank (Y_1) by an average of 0.00020. The increase is due to the bank's currency margin.

8. An increase in the net stable funding ratio (X_8) by 1, with other factors remaining unchanged, leads to a decrease in the return on assets of the bank (Y_1) by an average of 0.00068. The balance of cash flows in time and volume is achieved by reducing the bank's profit.

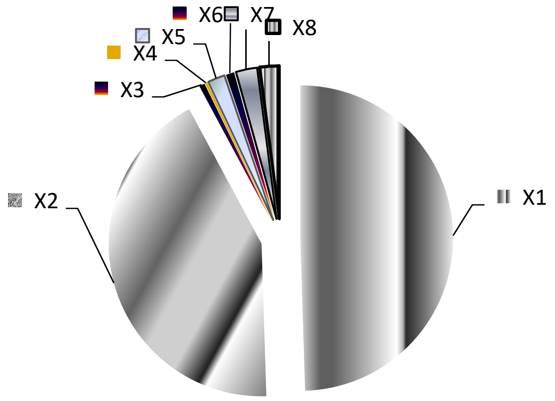


Fig. 2. The importance of factors on the return on assets of Ukrainian banks as at 2023.01.06 (source: authors on the base of [23] data)

The elasticity coefficients show that capital ratios have the greatest impact on the return on assets of banks, followed by liquidity ratios, and credit and currency risks have the least impact.

In model (1), the ranking of the X-factors by the strength of their impact on the performance indicator (Y_1) is as follows:

$X_1 > X_2 > X_7 > X_8 > X_5 > X_6 > X_3 > X_4$ (Figure 2).

As Figure 2 shows, capital ratios have the most significant impact on banks return on assets. The higher the regulatory capital adequacy, the higher the return on assets of banks. At the same time, the expansion of capitalisation deprives the bank of a portion of its profits for reinvestment. Therefore, the return on assets decreases as reserve capital increases. Banking risks (credit, currency, liquidity) have a significantly lower impact compared to capitalisation ratios. This is due to the peculiarities of Ukrainian banking management in the context of increased instability and external threats. This means that risks are significantly limited by the tools used to manage banking operations. For example, restricting lending, creating liquidity reserves in the form of profitable and risk-free government debt securities.

C. Model (2)

The economic-mathematical model (2) of efficiency of banks in Ukraine has the form:

$$Y_2 = 8.91289 - 0.02378 \cdot X_1 - 0.10327 \cdot X_3 - 0.00555 \cdot X_4 - \\ - 0.00001 \cdot X_6 + 0.00283 \cdot X_8 \qquad (2)$$

where Y_2 is the yield on the bank's loan portfolio;

X_1 – bank regulatory capital adequacy ratio;

X_3 – maximum exposure to credit risk per counterparty;

X_4 – risk ratio of the total long open currency position;

X_6 – liquidity coverage ratio for all currencies;

X_8 – net stable funding ratio ratio.

According to Fisher's criterion, it can be argued that model (2) is reliable with a probability of 0.74. This is proved by fulfilling the condition: the actual value of the Fisher criterion of 1.34379 exceeds its tabulated value at the level since the actual value is 1.34045.

First constraint for Y_2 = 6.34729. It is calculated based on the critical thresholds for the X-factors through the model (2): X_1=10%, X_3=25%, X_4=5%, X_6=100%, X_8=100%.

The second constraint for Y_2 = 8.91289. It is equal to the free term of model (2) and means what Y_2 will be if all X-factors are equal to zero.

D. Characteristics of the model (2)

Parameters of the model (2) are shown in the Table II.

TABLE II. CHARACTERISTIC OF THE ECONOMIC-MATHEMATICAL MODEL (2)

Factor	**Weight values of the model coefficients**	**Standard errors of the coefficients of the variables**	**T-test for the model coefficients**	**Elasticity coefficients**
X_1	-0.02378	0.01104	-2.15413	-0.18557
X_3	-0.10327	0.06691	-1.54360	-0.24751
X_4	-0.00555	0.01194	-0.46499	-0.00630
X_6	-0.00001	0.00010	-0.14468	-0.00239
X_8	0.00283	0.00285	0.99292	0.11117

[b.] calculated by the authors

Let us describe the parameters of model (2).

With an increase in the regulatory capital adequacy ratio (X_1) by 1, with other factors remaining constant, the level of profitability of the bank's loan portfolio (Y_2) decreases by 0.02378 on average.

An increase in the maximum credit risk exposure per counterparty (X_3) by 1, with other factors remaining unchanged, leads to a decrease in the level of profitability of the bank's loan portfolio (Y_2) by 0.10327 on average due to the creation of provisions.

An increase in the risk ratio of the total long open currency position (X_4) by 1, with other factors remaining constant, leads to a decrease in the level of profitability of the bank's loan portfolio (Y_2) by an average of 0.00555.

An increase in the liquidity coverage ratio for all currencies (X_6) by 1, with other factors remaining constant, leads to a decrease in the level of profitability of the bank's loan portfolio (Y_2) by an average of 0.00001.

An increase in the net stable funding ratio (X_8) by 1, with other factors remaining constant, leads to an increase in the level of profitability of the bank's loan portfolio (Y_2) by an average of 0.00283.

According to the elasticity coefficients, it can be seen that credit risk and bank capitalisation have the greatest impact on the level of profitability of the bank's loan portfolio. In model (2), the ranking of X-factors by the strength of their influence on the performance indicator (Y2) is as follows: $X_3 > X_1 > X_8 > X_4 > X_6$ (Figure 3).

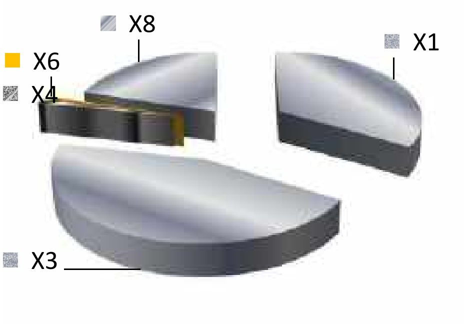


Fig. 3. The strength of influence of factors on the level of profitability of the loan portfolio of Ukrainian banks as at 2023.01.06 (source: authors on the base of [23] data)

Figure 3 shows that credit risk has the greatest impact on the profitability of banks' loan portfolios. The higher the assumed credit risk, the lower the profitability of the loan portfolio. The second largest factor is regulatory capital adequacy. The higher the level of security for the bank's solvency, the lower the efficiency of its loan portfolio. The third place was taken by the net stable funding ratio. Ensuring that pea flows are balanced over time and in terms of volume contributes to the profitability of the bank's loan portfolio. The last places with the least influence are shared by currency risk and liquidity risk. Exchange rate fluctuations cause negative changes in the profitability of the loan portfolio. Deterioration in the liquidity of banks' assets leads to a decline in loan portfolio profitability.

E. Model (3)

Mathematical model of multi-criteria choice in the field of bank lending.

Another rather important aspect of the activity of banking institutions and the entire banking sector in general is lending. As is known, with the development of market relations, the process of lending by banks to enterprises is associated with numerous risk factors that can cause non-repayment of the loan on time. When analyzing the borrower's creditworthiness, the possibility of timely and full repayment of the loan debt is determined; the degree of risk that the bank is willing to take; the amount of credit that can be granted in a specific situation; terms of granting a loan.

In today's conditions, the analysis of creditworthiness is connected not only with the assessment of the client's solvency on a certain date, but also with the identification of the best borrowers, forecasting their financial stability in the future, accounting for possible risks in credit transactions and a number of other factors. Conducting such a comprehensive analysis allows a Ukrainian bank to more effectively manage credit resources and obtain profit in modern conditions.

We will build a mathematical model based on the use of the theory of fuzzy sets and multi-criteria optimization in the field of lending, which allows us to increase the validity of the decisions made and ensure the selection of the most rational options from the set of admissible ones.

Let the institutions turn to some bank with a request to provide them with loans. Since the bank's resources are limited, it faces the task of choosing the best institution based on a set of quality criteria. In the considered task, institutions can be marked as alternatives, from which the best choice should be made. We will denote the alternatives.

Data from their accounting statements are used to assess the creditworthiness of borrower institutions:
cash (C), short-term financial investments (FV),
accounts receivable (AR), stocks and costs (SC),
equity capital (EC), short-term liabilities (STL),
balance sheet summary (BS), gross turnover (GT), profit (P),
on the basis of which coefficients characterizing the creditworthiness of borrowers are calculated:
coefficient of absolute liquidity (f_1),
intermediate coverage ratio (f_2),
total coverage ratio (f_3),
coefficient of financial independence (f_4),
product profitability ratio (f_5).

The general formulation of the problem of determining the combination of alternatives with maximum efficiency (or efficiency per unit of the required resource) consists in determining combinations of alternatives that satisfy the specified objective functions: coefficients characterizing the creditworthiness of borrowers f_1, f_2, f_3, f_4, f_5.

To build a mathematical model, consider the concept of a fuzzy multiset and the multiplicity of the elements of this set according to [22, 24].

On the basis of the given characteristics, we can calculate the values of the quality criteria for the considered institutions, give the normative values of the criteria. When analyzing the estimated and normative values of the criteria, it may be that all institutions can apply for a loan. After that, we can apply decision theory and fuzzy set theory.

Processing of input information using the mathematical apparatus of fuzzy set theory is carried out in three stages.

1. Construction of ownership functions corresponding to the concepts of "best absolute liquidity ratio", "desired intermediate coverage ratio", "best profitability ratio", etc. The construction of such functions is carried out by experts who have knowledge in the field of lending to enterprises of various functional purposes.

2. Specific values of membership functions according to quality criteria f_1 , f_2 , f_3 , f_4 , f_5 corresponding to the considered alternatives are determined. Fuzzy sets for the five considered criteria comprising the alternatives under analysis.

3. The available information is collated in order to identify the best alternative. The set of optimal alternatives is determined by crossing the fuzzy sets containing the evaluations of the alternatives according to the selection criteria.

We will consider the alternatives that maximize the vector criterion and the functions that have the maximum value, belonging to the set, to be optimal. The operation of the intersection of fuzzy sets corresponds to the selection of the minimum value for the *i*-th alternative.

CONCLUSIONS

The work deals with a rather important topic of data analysis, in particular, the analysis of indicators of the bank's activity.

Three economic-mathematical models were built and statistical analysis of data was made based on decision-making theory and correlation-regression analysis.
Calculations were made on the basis of the latest statistical indicators and summary data. These models can be used for implementation in banking institutions.

REFERENCES

[1] Ashraf Bany Mohammad, Manaf Al-Okaily, Mohammad Al-Majali, Ra'ed Masa'deh, Business Intelligence and Analytics (BIA) Usage in the Banking Industry Sector: An Application of the TOE Framework, J. Open Innov. Technol. Mark. Complex. 2022, 8, 189. https://doi.org/10.3390/joitmc8040189

[2] Al-Okaily, M.; Al-Okaily, A. An Empirical Assessment of Enterprise Information Systems Success in a Developing Country: The Jordanian Experience. TQM J. 2022

[3] Nithya, N.; Kiruthika, R. Impact of Business Intelligence Adoption on performance of banks: A conceptual framework. J. Ambient. Intell. Humaniz. Comput. 2021, 12, 3139–3150.

[4] Ajah, I.A.; Nweke, H.F. Big data and business analytics: Trends, platforms, success factors and applications. Big Data Cogn. Comput. 2019, 3, 32.

[5] V. Yu. Kochorba and O.V. Manets, "The efficiency of banks with foreign capital on the basis of econometric modelling", International scientific journal "Internauka". Series: Economic Sciences, vol. 5, issue 1, 2021, pp. 29–37.

[6] I. Kramar, H. Tsikh, I. Nahorniak and L. Pokryshka, "Improvement of strategic management of the bank on the basis of econometric modelling of its efficiency", Socio-economic problems and the state, vol. 2, 2021, pp. 457–464.

[7] O. Badunenko, S. C. Kumbhakar and A. Lozano-Vivas, "Achieving a sustainable cost-efficient business model in banking: The case of European commercial banks", European Journal of Operational Research, vol. 293, issue 2, 2021, pp.773-785.

[8] I. H. Abernikhina, "Improving the model of forecasting bank profit using correlation and regression analysis", Business Navigator, vol. 2, 2020, pp. 79-86.

[9] V. M. Domrachev and V.V. Tretynyk, "Modelling the impact of the economic environment on the performance indicators of a Ukrainian bank", Economics and management, vol. 1, 2020, pp. 118-127.

[10] D. Corbae and P. D'Erasmo, "Capital requirements in a quantitative model of banking industry dynamics", National Bureau of Economic Research. Cambridge, vol. wp25424, 68 p. January 2019.

[11] D. Corbae and P. D'Erasmo, "Capital buffers in a quantitative model of banking industry dynamics", Econometrica, vol. 89, issue 6, 2021, pp. 2975-3023.

[12] L. O. Prymostka and N.S. Sokolovska, "Measurement (assessment) and modelling of a bank's operational risk", Business Inform, vol. 11, 2021, pp. 144–153.

[13] A. V. Oliinyk, "Assessment of losses from credit risk and modelling of bank liquidity risk", Bulletin of Khmelnytsky National University. Economic Sciences, vol. 3, 2021, pp. 323–332.

[14] K. G. M. Karvana, S. Yazid, A. Syalim and P. Mursanto, "Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry," 2019 International Workshop on Big Data and Information Security (IWBIS), Bali, Indonesia, pp. 33-38, October 2019.

[15] O. M. Vasyliev, "Scoring modelling based on neural networks for determining the rating of a bank borrower", Ekonomika Ukrainy, vol. 10, 2020, pp. 54-62.

[16] I. Yanenkova, Yu. Nehoda, S. Drobyazko, A. Zavhorodnii and L. Berezovska, "Modeling of Bank Credit Risk Management Using the Cost Risk Model", Risk and Financial Management, vol.14, issue 5, 211, 2021, pp. 1-15.

[17] O. V. Kuzmenko, H. M. Yarovenko and L. O. Skrynka, "Analysis of mathematical models of counteracting bank cyber fraud", Bulletin of Sumy State University. Series: Economics, vol. 2, 2022, pp.111-120.

[18] O. Kuzmenko, T. Dotsenko and O. Kushnerov "Assessment of the risk of using banks for the purpose of legalisation of criminal proceeds based on gravity modelling", Problems and Prospects of Economics and Management, vol. 1, 2020, pp. 205-219.

[19] Cerneviciene J and Kabašinskas A Review of Multi-Criteria Decision-Making Methods in Finance Using Explainable Artificial Intelligence. Front. Artif. Intell., 2022 5:827584. doi: 10.3389/frai.2022.827584

[20] L. Koliechkina, O. Pichugina and O. Dvirna, "Horizontal Method Application to Multiobjective Combinatorial Optimization over Permutations," 2022 IEEE 3rd International Conference on System Analysis & Intelligent Computing (SAIC), Kyiv, Ukraine, 2022, pp. 1-5, doi: 10.1109/SAIC57818.2022.9923018.

[21] Pichugina, L. Koliechkina and T. Chilikina, "Multicriteria Combinatorial Optimization Model of an Infocommunication System," 2021 IEEE 8th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T), Kharkiv, Ukraine, 2021, pp. 13-16, doi: 10.1109/PICST54195.2021.9772124.

[22] Semenova N.V., Kolechkina L.M.Vector problems of discrete optimization on combinatorial sets: research and solution methods: Monograph. - Kyiv: Naukova dumka, 2009. - 266 p.

[23] The National Bank of Ukraine. [Online]. Available: https://bank.gov.ua/ua/statistic.

[24] Koliechkina, L.N., Dvirna, O.A. & Khovben, S.V. A Two-Step Method for Solving Vector Optimization Problems on Permutation Configuration. Cybern Syst Anal 57, 442–454 (2021). https://doi.org/10.1007/s10559-021-00369-3

Modeling the Dynamics of the EU Stock Indices Based on the Analysis of Structural Market Data

Anastasiia Shenderova
Master Student
Department of World Economy and International Economic Relations
Odesa I. I. Mechnikov National University
Odesa, Ukraine
0009-0001-0303-7197

Sergiy Yakubovskiy
Doctor of Economics, Professor
Department of World Economy and International Economic Relations
Odesa I. I. Mechnikov National University
Odesa, Ukraine
0000-0002-1193-0241

***Abstract* — The last several years were quite challenging for the whole world. People, businesses, and governments must adapt to new realities and find new ways to survive. The shock of the world pandemic brought uncertainty to quotidian life, then the Russian-Ukrainian war forced people to accept new obstacles. Meanwhile, governments must look for paths to keep calm, not only nations but also make everything possible to keep low economic indicators such as interest rate, inflation, unemployment, etc. within all these periods. In this study, the statistical importance of dependence of earlier mentioned and statistical metrics including exchange rate, current account, inflow and outflow of portfolio investment on the price of the stock indexes of Germany, France, Italy, and Greece is examined. These EU countries have been strategically selected due to their pivotal roles in the European economic landscape, coupled with their acute susceptibility to instability-triggered fluctuations. For instance, the resonance of social mass protests in France, and the intricate interplay of national debt increasing in Italy and Greece, make them particularly interesting cases for this analysis. By examining the interplay between economic indicators and stock index movements in these nations, this study sheds light on the relationship between macroeconomic variables and financial market performance. Through rigorous statistical analysis, the aim of this research is to uncover potential patterns and dependencies that might offer valuable insights into the ways in which economic dynamics interact with stock market trends in times of upheaval. Ultimately, this research aspires to contribute to a deeper understanding of the mechanisms underpinning the equilibrium between economic fundamentals and financial market behavior during tumultuous periods.**

Keywords* — *stock index, DAX, CAC40, MIB, ATHEX, correlation analysis, regression analysis

I. Introduction

The global landscape is characterized by rapid and constant change, underscoring the paramount importance of staying well-informed and attuned to significant developments across political, socio-economic, and environmental spheres. Over the past few years, we've witnessed a series of transformative events. A global pandemic struck, precipitating an economic crisis of substantial magnitude. Additionally, in 2022 population witnessed a full-scale conflict erupting on Ukrainian soil, which not only exacerbated the ongoing economic crisis but also catalyzed a social upheaval.

These historical incidents have exerted a profound influence on the trajectory of the global economy as well as the individual economies of nations. Europe emerged as a region significantly impacted by these occurrences. This prompts our academic curiosity to delve into a comprehensive study of recent stock market dynamics within the principal EU countries. Our selection of stock markets includes the German market, chosen due to the resilience of its national economy; the French market, selected in light of prolonged societal discord within the influential nation; the Greek market, in consideration of its economy's susceptibility to past crises; and the Italian market, due to the looming specter of potential debt crises.

By examining the trends and patterns within these chosen markets, further research will be built on gaining a deeper understanding of how these economies have responded to and navigated through the intricate web of challenges posed by recent events.

II. Analysis of Macroeconomic Indicators Affecting the Stock Indices of the EU Countries

Commencing this investigation with an overview of the fluctuations in chosen stock markets holds significance, offering valuable insights to guide financial choices, evaluate potential risks and opportunities, and analyze the prevailing condition of both the economy and the market.

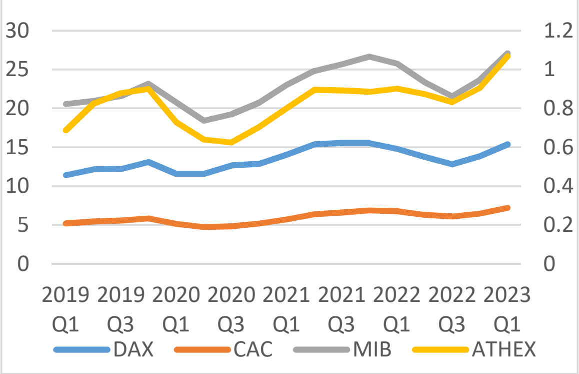


Fig. 1 Dynamics of DAX, CAC, MIB and ATHEX indices (right axis) during 2019 Q1 – 20231 Q1.

Source: compiled by the authors based on [1], [2], [3], [4]

Examining Fig. (1) allows to deduce the overall trend of all indices as moving in an upward direction. This trend is logical, as over the long term, prices tend to rise. Nevertheless, the chosen indices can be categorized into two distinct groups based on their patterns. The first group comprises the German index (DAX) and the French index (CAC), while the second group consists of the Italian index

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

(MIB) and the Greek index (ATHEX). Analyzing the first pair of indices, we can observe that the French stock index mimics the movements of the German index, albeit with a certain time lag. Notably, the volatility of the CAC is lower in comparison to the DAX's volatility.

Similar dynamics can be seen in the case of the Italian and Greek indices, where the former experiences a delay in its movements when compared to the latter. These indices clearly illustrate the decline of the European stock market during moments of global phenomena and crises. Such instances include the periods of January to March 2020, marked by the initial spread of the global Covid-19 pandemic, encompassing extensive restrictions and lockdowns. Additionally, the months of September to October 2021 (extending to December 2021 - January 2022 in Italy) are notable, corresponding to the heightened escalation of the conflict within Ukraine's territory. This conflict ultimately culminated in a full-fledged modern war, which persists to the present day.

According to the results of the research by W. Zhao, it is worth noting that six months before the Russian-Ukrainian conflict (August 2021) until now (August 2022), the world was still in the midst of a pandemic, and the number of infections continued to rise. This has led to concerns about the market economy and the potential impact of the continued rise in infections on the stock market. This factor also encouraged people to sell their stocks to prevent possible losses in the event of a stock market crash. Research shows that during the Russian-Ukrainian conflict, crude oil prices have a strong impact on stock markets, so stock investors are advised to be cautious and carefully analyze the situation during the military conflict before making investment decisions [5].

TABLE I. CORRELATION MATRIX OF DAX, CAC, MIB AND ATHEX INDICES

	DAX	CAC	MIB	ATHEX
DAX	1			
CAC	0.886493	1		
MIB	0.905164	0.93387	1	
ATHEX	0.734556	0.86339	0.834581	1

Source: compiled by the authors based on [1], [2], [3], [4]

From the correlation matrix Tab. (1), we can conclude that the DAX, CAC, MIB, and ATHEX indices show a strong relationship, and their movement is often parallel to each other in the presented time range. This may indicate that macroeconomic or global events may have a similar impact on these markets.

The monetary policy of the ECB has a profound effect on the financial markets of all the European Monetary Union countries and affects their securities and investment opportunities. To the extent that all states are part of the European Monetary Union, the European Central Bank (ECB) 's actions in monetary policy have a similar effect on the securities markets of these countries. This means that measures taken by the ECB, such as changes in refinancing rates, asset purchase programs, or other monetary measures, have a joint effect on the financial situation of member countries. This similarity in reactions to monetary and credit measures of the ECB can significantly affect the circulation of securities in financial markets. In such a circumstance, changes in the ECB's policy may cause similar reactions in the securities markets of all member countries, as a result of which the prices of shares, bonds, and other securities may change in accordance with the actions of the ECB [6].

Another possible reason for such a strong relationship is the European economic connection since all four countries are in relative geographical proximity. The strong positive correlation between these indices may reflect general economic trends and developments in Europe. European economies can be interdependent, and macroeconomic events such as changes in regional politics, trade relations, or world conditions can have a similar effect on indices. But the very concept of global events occupies not the last place in the rating of influence on the stock markets not only in Europe but also in the world. The relationship of indices can reflect their reaction to global events such as global economic crises, geopolitical tensions, or even changes in world trade dynamics. If one region experiences strong influences as a result of global events, this can be reflected in other regions through interconnections in the global economy.

Based on the findings of the study conducted by S. Liu, changes in one of the endogenous variables cause fluctuations in other variables. In other words, the volatility of the stock market in each country affects other countries to a different extent during special events, which gives us an idea to improve the current situation in the financial markets of each country. Therefore, governments should take into account the situation in the stock markets of other countries in order to take effective measures to prevent the impact of Covid-19 on the stock markets [7].

It is important to emphasize that although correlation reveals relationships between indices, it does not necessarily indicate causality. Correlation can be the result of a variety of factors, and a more accurate understanding of these relationships may require additional analysis and study of specific events and trends in each time period. As already mentioned below, the regression model of the impact of the main macroeconomic indicators on the indices of these countries will be considered. The stock market performance is vigorously influenced by distinct macroeconomic indicators and this varies depending on the country. So, the following general model for countries is obtained:

$$A = R*\beta1 + B*\beta2 + G*\beta3 + I*\beta4 + E*\beta5 + C*\beta6 + N*\beta7 + O*\beta8, \quad (1)$$

where A – country index price, R – unemployment rate, B – government debt, G – country's real GDP growth rate, I – interest rate (EMU convergence criterion series), E – exchange rate, C – current account, N – inflow of portfolio investments, O - outflow of portfolio investments. It should be noted that A is a dependent variable, all other indicators are independent variables.

In this research, the main sources of statistical data were information from the statistical databases of world organizations, such as the World Bank, the European Central Bank (ECB), the Organization for Economic Cooperation and Development (OECD), and others. Due to the data limitation as the lack of certain indicators with monthly frequency, quarterly data were used. The timeframe was taken from 2016 Q1 to last available (2023 Q1). The data for the indices were taken based on the calculation of the average value during the entire quarter in order to reflect all possible fluctuations in the share price of the index. For the next stage of the research, the Stata program was used, which made it possible to perform regression analysis quickly and without unnecessary complications.

A comprehensive regression model for countries was presented earlier. Subsequently, an individual analysis was conducted for each country. Within each regression model utilizing the previously mentioned variables, statistically insignificant factors were identified. These factors were excluded as they do not impact the dependent variable and can lead to an elevation in the variance of OLS estimations.

TABLE II. REGRESSION ANALYSIS OF GERMAN STOCK INDEX DAX ON MAIN MACROECONOMIC INDICATORS

Variable	Coefficient	Std. error	t-statistics	Probability
R	-6.996	1.784	-3.92	0.001
B	0.408	0.118	3.46	0.002
E	11.610	4.992	2.33	0.028
_cons	-11.006	8.238	-1.34	0.194

Source: compiled by the authors

According to the results of the statistical analysis of the regression model in Tab. (2), it was found that the influence of the general group of independent variables on the dependent variable is statistically significant. This is confirmed by the value of the F-statistic (5.44) and the corresponding p-value (0.0051), which is less than the specified level of significance.

The unemployment rate revealed a negative and statistically significant regression coefficient, indicating an indirect relationship between this variable and the dependent variable: as unemployment decreases, the price of the national index is expected to increase. Additionally, Germany's budget deficit and the exchange rate of the currency against the dollar show statistically significant positive coefficients.

R-squared is 0.3950, which can be interpreted as the fact that approximately 39.50% of the variation in the dependent variable can be explained by changes in the independent variables that are considered in the model.

The analysis of the results of the statistical study of the regression model in Tab. (3) confirms the important influence of the general group of independent variables on the dependent variable. This is supported by the F-statistic value (16.73) and the corresponding p-value (0.0000), which does not exceed the accepted level of significance.

TABLE III. REGRESSION ANALYSIS OF FRENCH STOCK INDEX CAC40 ON MAIN MACROECONOMIC INDICATORS

Variable	Coefficient	Std. error	t-statistics	Probability
R	-0.984	0.141	-6.96	0.000
G	0.040	0.020	2.00	0.057
I	0.346	0.113	3.06	0.005
C	0.021	0.008	2.48	0.020
_cons	10.673	0.776	13.74	0.000

Source: compiled by the authors

The regression coefficients for the unemployment rate, the interest rate on long-term government bonds, and the national current account show statistical significance with high p-values. Meanwhile, for the growth of the real GDP of France, a statistically significant regression coefficient is found only at the level of significance at the level of 10%. It is important to note that only the unemployment rate in France has a negative regression coefficient, while all others have positive ones: as the unemployment rate increases, the price of the SAC index decreases, while the other independent variables have the opposite effect.

The coefficient of determination (R-squared) reaches a value of 0.7361, indicating about 73.61% of the variation of the dependent variable can be explained by changes in the independent variables within this model.

TABLE IV. REGRESSION ANALYSIS OF ITALIAN STOCK INDEX MIB ON MAIN MACROECONOMIC INDICATORS

Variable	Coefficient	Std. error	t-statistics	Probability
R	-3.668	0.822	-4.46	0.000
B	-0.142	0.073	-1.94	0.065
G	0.341	0.115	2.95	0.007
E	27.905	8.790	3.17	0.004
C	-0.084	0.046	-1.80	0.085
N	0.049	0.027	1.81	0.084
_cons	30.778	13.109	2.35	0.028

Source: compiled by the authors

The results of the statistical analysis of Tab. (4) indicate a significant influence of the general group of independent variables on the dependent variable. This is confirmed by the low value of the F-statistic (6.76) and the corresponding p-value (0.0004), which does not exceed the accepted level of significance.

The regression coefficients for Italy's unemployment rate are statistically significant and negative, causing the price of the selected index to decrease as this independent variable increases. Real GDP growth and the euro-to-dollar exchange rate have positive statistically significant regression coefficients. Only if the level of significance is chosen at the level of 10%, significant coefficients for the country's budget deficit, the national current account, and the inflow of portfolio investments to Italy are revealed. At the same time, the first two coefficients are negative, and the last one is positive.

R-squared is 0.6485, which indicates the possibility of explaining about 64.85% of the variation of the dependent variable with the help of the influence of the independent variables included in the model.

TABLE V. REGRESSION ANALYSIS OF GREEK STOCK INDEX ATHEX ON MAIN MACROECONOMIC INDICATORS

Variable	Coefficient	Std. error	t-statistics	Probability
R	-0.055	0.006	-8.61	0.000
B	-0.006	0.001	-3.99	0.001
E	1.216	0.339	3.59	0.001
_cons	1.260	0.275	4.57	0.000

Source: compiled by the authors

The results of the statistical analysis of the regression model in Tab. (5) showed that the overall effect of the independent variables on the dependent variable is statistically significant. This is confirmed by the F-statistic value (24.89) and the corresponding p-value (0.0000), which is less than the chosen significance level of 5%.

The level of unemployment and the budget deficit of Greece also have negative and statistically significant regression coefficients, which state a fall in the share price of the selected index when the mentioned independent ones increase.

The coefficient of determination (R-squared) is 0.7492, which means that about 74.92% of the variation of the dependent variable can be explained by changes in the independent variables in the model.

Based on the analysis of the presented models, it can be concluded that their quality leaves room for improvement. Therefore, in subsequent research, a more detailed examination of these models and the implementation of relevant modifications to increase the coefficient of determination R2 should be considered. Possible methods to enhance model quality may include expanding the sample size, exploring alternative independent variables etc.

Conclusion

The rapidly evolving global landscape underscores the importance of staying informed about developments across political, socio-economic, and environmental domains. Recent transformative events, including a global pandemic, economic crisis, and conflict in Ukraine, have significantly impacted global and national economies.

The chosen indices generally show an upward trend, with distinct patterns grouping them into German/French and Italian/Greek clusters. These indices illustrate the impact of global events on the European stock market.

The correlation matrix emphasizes a strong connection among the DAX, CAC, MIB, and ATHEX indices. The ECB's monetary policy significantly influences EU financial markets. Individual regression analyses identified significant factors, yielding insights into unemployment's inverse relationship with stock indices.

It's also worth noting a couple of causal relationships that are also reflected in the regression analysis. First, the devaluation/revaluation of the national currency leads to an increase/decrease in the price of the stock index. This result is also evident in three out of four of the presented regression analysis tables, where the exchange rate demonstrates significance. Second, there is an inverse relationship between unemployment and the value of stock index prices, which can be explained through the Phillips Curve since inflation and unemployment are inversely related, and rising prices also lead to higher stock prices.

It can be inferred that during periods of elevated unemployment, regulatory authorities across all examined countries, including the ECB, strive to invigorate economic activity through monetary easing measures, subsequently resulting in an upswing in stock indices. Consequently, a suggestion can be offered to participants in the stock market: during the height of the unemployment rate and following its initiation of decline, they should consider initiating stock market investments. This is because, after the commencement of the unemployment rate reduction, the upturn in stock indices is likely to ensue. Despite limitations, the study contributes to understanding market fluctuations, emphasizing the interconnectedness of global markets and the importance of considering other nations' stock markets in strategic planning.

References

[1] Börse Frankfurt - Frankfurt Stock Exchange. [Online] Available: https://www.boerse-frankfurt.de/en [Accessed Aug.15, 2023].

[2] Euronext Paris. [Online] Available: https://www.euronext.com/en/markets/paris [Accessed Aug.15, 2023]

[3] Euronext Milan. [Online] Available: https://www.borsaitaliana.it/homepage/homepage.htm [Accessed Aug.15, 2023].

[4] ATHEXGROUP. [Online] Available: https://www.athexgroup.gr/web/guest/home [Accessed Aug.15, 2023]

[5] W. Zhao "Changes in Crude Oil Prices under the Russia-Ukraine Conflict". Highlights in Business, Economics and Management, vol.8, pp. 337–347, 2023.

[6] V. Tuliakov, H. Alekseievska i S. Yakubovsky, "Legal and economic aspects of monetary regulation of the european system of central banks", InterEULawEast, vol.8, n. 2, pp. 79-96, 2021.

[7] S. Liu (2023) COVID-19 pandemic: measuring stock indices correlation between different countries. Highlights in Business, Economics and Management, vol.10, pp. 65-71, 2023.

[8] Eurostat Data (2023). [Online] Available: https://ec.europa.eu/eurostat [Accessed Aug.15, 2023].

[9] OECD Data (2023). [Online] Available: https://data.oecd.org/ [Accessed Aug.15, 2023].

Density-Based Outlier Detection: Supervised Approach Based on Virtual Points

Olena Domanska
Ivan Franko National University of Lviv, Avenga
Lviv, Ukraine
olena.domanska@lnu.edu.ua

Valerii Martsyshyn
Ivan Franko National University of Lviv, Avenga
Lviv, Ukraine
valerii.martsyshyn@lnu.edu.ua

Oleh Buhrii
Ivan Franko National University of Lviv
Lviv, Ukraine
oleh.buhrii@lnu.edu.ua

***Abstract* — Outlier detection plays an important role in a wide variety of domains and has many applications. The examples are fraud detection, anomaly detection in industrial systems, healthcare, cybersecurity, environmental monitoring and marketing. The aim of outlier detection is to identify and analyze data points or observations that differ significantly from the rest of the data that may indicate potential fraud, disease outbreaks, or equipment failures. The unsupervised approach to outlier detection is often preferred because it does not require labeled data, meaning that the algorithm can identify outliers in a dataset without having to know beforehand which observations are anomalous. This is particularly useful when there is limited prior knowledge about the data, or when the anomalous observations are rare or unknown. We propose a novel approach to outlier detection which doesn't require labeled data but is based on the supervised learning technique similar to decision tree construction. Our method overcomes many of the limitations of traditional outlier detection techniques by partitioning the space into cluster and empty regions using a decision tree. We accomplish this by introducing virtual data points and modifying the decision tree algorithm accordingly. The results of our experiments on synthetic and real-world datasets show that the method is both highly efficient and scalable.**

Keywords — outlier detection, anomaly detection, clustering, decision tree construction.

I. Introduction

A. Anomaly detection

Outlier detection is an important task in data analysis and decision making because it helps identify and analyze data points or observations that deviate significantly from the rest of the data. These outliers may indicate potential fraud, equipment failures, disease outbreaks, or other anomalies that require attention or corrective action.

Outlier detection and has many applications. The examples are fraud detection, anomaly detection in industrial systems, healthcare, cybersecurity, environmental monitoring and marketing. Outlier detection is used extensively in the banking and financial industry to detect fraudulent transactions. This helps to prevent financial losses due to fraudulent activities. Speaking about industrial systems, outlier detection is used to detect anomalies in the performance of machinery and equipment. This helps to identify and prevent potential equipment failures, reducing downtime and maintenance costs. In healthcare outlier detection helps to identify patients who are at high risk of developing certain conditions, such as diabetes or heart disease, enabling the possibility to provide early intervention and preventive care, improving patient outcomes. In cybersecurity outlier detection is used to detect and prevent cyber-attacks, such as intrusion detection, malware detection, and network traffic analysis. Besides, outlier detection can be successively leveraged in environmental monitoring to identify unusual patterns in the behavior of ecosystems, such as changes in water quality or air pollution levels. Marketing can benefit from outlier detection as well to identify unusual patterns in consumer behavior, such as sudden changes in purchasing habits, which can be used to develop targeted marketing campaigns.

Outlier detection is a widely researched topic in statistics, machine learning, and data mining. Common approaches in the literature can be classified into the following groups: density-based methods, distance-based methods, model-based methods and ensemble methods. Density-based methods identify outliers as data points with low local densities, which are far away from the clusters or have too few neighboring points. Examples of density-based outlier detection methods include DBSCAN [1] and OPTICS [2]. Distance-based methods identify outliers as data points that are far away from the centroids of their assigned clusters or from other data points. Examples of distance-based outlier detection methods include k-NN [3] and LOF [4]. Model-based methods assume that the data is generated from a specific statistical model, and identify outliers as data points that do not fit this model well. Examples of model-based outlier detection methods include Gaussian mixture models and one-class SVM [5]. Ensemble methods combine multiple outlier detection methods to improve the accuracy and robustness of the results. Examples of ensemble outlier detection methods include Isolation Forest [6] and Local Outlier Factor Ensemble. Each method has its own strengths and weaknesses, and the choice of method depends on the specific characteristics of the dataset and the goals of the analysis.

B. Outlier Detection as an Initial Step in Data Analysis

Although outlier detection is an important standalone task, it is often used as a preliminary step in various problems in data analysis, machine learning, and statistics. Outliers can be caused by various reasons such as measurement errors, data entry errors, or rare events. The presence of outliers in a dataset can affect the accuracy and reliability of statistical models and machine learning algorithms. Therefore, detecting and handling outliers is an important step in many data analysis tasks.

For instance, outlier detection could be very helpful as a preliminary step in clustering, as it helps identify and remove

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

data points that do not belong to any cluster or that belong to a different cluster than the one they were assigned to.

Clustering is one of the most important tasks in Statistics and Data Science. It aims to extract valuable information from data and to group the records by similarities which are not recognizable by humans due to their complexity or amounts of data. Information extracted through clustering if treated in the right way can lead businesses to advantage on the market. If not, or if a clustering algorithm performs poorly, it may result in building inefficient models and cause drops in revenue. That's why it is so important to have an algorithm which works the best and gives understanding for humans on similarities inside clusters.

Our paper presents a new approach to outlier detection which we call AnomalyTree. It utilizes a supervised learning method known as decision tree construction even though the labels for outliers are not known. This novel technique draws inspiration from CLTree (Clustering based on Decision Trees) [7] and differs significantly from existing methods, offering several unique advantages.

The Source Code is available on GitHub: https://github.com/DomanskaOlena/density-based-outlier-detection and contains the implementation of the proposed method in Python as well as the comparison of our method's performance with well-known anomaly detection methods.

II. Approach Description

The focus of our research is outlier detection in a numerical space, characterized by each dimension or attribute having a bounded and ordered domain.

The methodology for outlier detection that we present in this paper relies on what is known as "virtual points", which were initially presented in [7]. Authors suggested an idea of a clustering algorithm based on adjusted decision trees called *CLTree*. The fundamental concept is to assign each data record or point in the dataset a class Y. Additionally, we assume that the data space comprises uniform distribution with another point type, referred to as non-existing points, which are assigned the class N. By incorporating the N points into the initial data space, we transform the task of segregating the data space into dense and sparse regions into a classification problem. Solving the problem entails the application of a decision tree algorithm.

To demonstrate the rationale behind the proposed method, we employ an illustration. Figure 1 (a) displays the 2D dataset in which two clusters (represented by green dots) and a set of outliers (represented by red dots) can be discerned. Following that, a set of N uniformly distributed points (represented in blue) is added to the data space. With this expanded dataset, a decision tree algorithm can be executed to obtain a partitioning of the space, as depicted in Figure 1(b). As a result of this process, the sparse regions containing the outliers are identified.

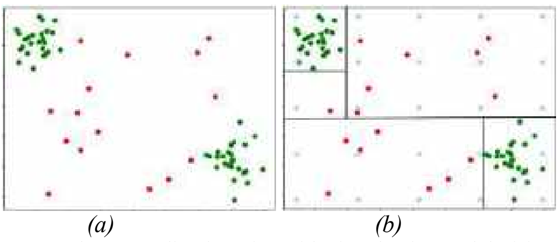


Fig. 1. AnomalyTree outlier detection with virtual points and decision trees

The AnomalyTree approach has many distinctive advantages compared to other methods. Firstly, the algorithm operates without any prior assumptions or input parameters. In contrast, many existing methods necessitate the user to determine the number of clusters required (such as GMM) and/or density thresholds (like DBSCAN), which can be challenging and arbitrary. Thus, the outliers identified may not adequately reflect the real arrangement of the data. Additionally, the technique can detect outliers both in the entire dimension space and in any subspaces, which is not always feasible. Frequently, algorithms that function efficiently in the high-dimensional space may not perform well in subspace, and vice versa. This method can be utilized for both cases, it can incorporate all dimensions or any subset thereof.

A. Detecting Sparse Regions by Introducing Virtual Points

Let's delve deeper into the concept of virtual points. As mentioned earlier, each data point in the original dataset is assigned a class label Y. Additionally, we introduce a set of "non-existing" N points that are uniformly distributed. It's important to note that these N points are not actually added physically to the original data; instead, their existence is assumed.

Now, the question arises, how many N points should be added? Initially, we add the same number of N points as there are Y points. However, during the tree construction process, each node receives a different number of N points based on a specific rule. If the number of N points passed down from the parent node to the current node is less than the count of Y points, we increase the number of N points in the current node to match the count of Y points. Conversely, if the inherited count of N points from the parent node is greater than or equal to the count of Y points, we use that inherited count.

This rule ensures that we increase the number of N points in a node if it has more inherited Y points than N points. This action is taken to avoid a situation where there might not be enough N points remaining after certain cuts or splits. Insufficient N points can pose challenges when performing further splits, even if those divisions are still necessary. Notably, the number of N points is not reduced if the current node is an N node, meaning it has more N points than Y points.

B. Modifications to the decision tree algorithm

At the heart of the proposed method is the decision tree construction algorithm. The decision tree construction algorithm typically follows the divide and conquer strategy, recursively dividing the data to create the tree. At each step, the algorithm selects the best cut to partition the data space in order to achieve purer regions. The information gain is commonly used as a criterion to determine the optimal cut. To calculate the information gain, we need to know the frequency or number of data points for each class on either side of a potential cut. Since we don't have N points, we have to compute them. Assuming the data points are uniformly distributed, we determine the number of N points on each side of the split proportionally to the area of the resulting regions. By utilizing these computed values, we can determine the information gain at each side.

Furthermore, we will evaluate the benefits of making cuts on both sides of the data points. In the traditional process of constructing decision trees, cuts are typically made on one side of the data points. However, it may result in inaccurate descriptions of the empty regions.

This issue leads to misidentifying normal data points as anomalies. To address this, we utilize a new approach that incorporates information gain while also looking ahead. The main concept is as follows: for each dimension (denoted as *i*), after identifying the initial cut using the gain criterion, we will search ahead (maximum of 2 steps) along each dimension to find a more optimal cut, denoted as c_i. This cut will help identify the best region, denoted as r_i, which is relatively empty measured by its relative density. The relative density of a region, *r*, is computed as *r.Y / r.N*, where *r.Y* represents the number of Y points and *r.N* represents the number of N points in region *r*. The cut c_i from dimension *i*, with the lowest relative density in its corresponding region r_i, will be selected as the most effective cut. This modified criterion aims to locate the least populated regions along each dimension. For more details please refer to [7].

III. Application to Anomaly Detection

The AnomalyTree algorithm detects anomalies by finding suitable cuts one dimension at a time, while many other anomaly detection algorithms take into account linear combinations of multiple features. Therefore, the proposed method works best for the problems where a significant deviation along one dimension would already signal an anomaly. For instance, in the problem of fraud detection in financial transactions, a significantly larger transaction amount or significantly more frequent transactions may already signal fraud, even if all other indicators remain on the same level. Moreover, since the AnomalyTree algorithm can find outliers in one dimension, it will also be able to cut points which are outliers in multiple dimensions.

Next we empirically evaluated the proposed method using synthetic datasets. Our experiments established the efficiency and accuracy of our method. It is capable of finding anomalies both in subspaces as well as in the full high dimensional space. Figure 2 shows the anomalies found in a two dimensional synthetic dataset. It might be difficult or even impossible to cut the anomalies using distance-based approaches in this case, but the AnomalyTree algorithm did a perfect job:

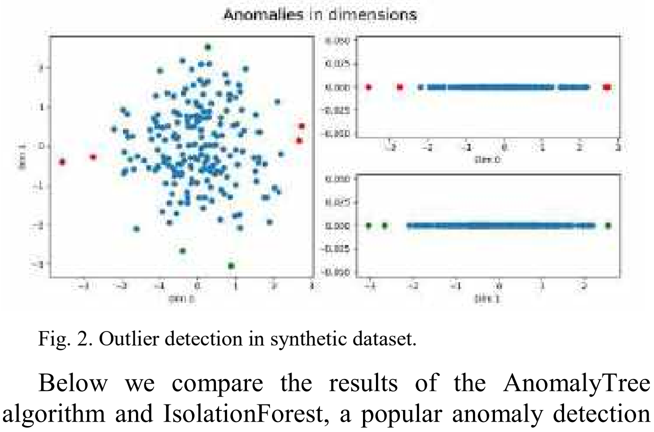


Fig. 2. Outlier detection in synthetic dataset.

Below we compare the results of the AnomalyTree algorithm and IsolationForest, a popular anomaly detection algorithm, on the synthetic dataset:

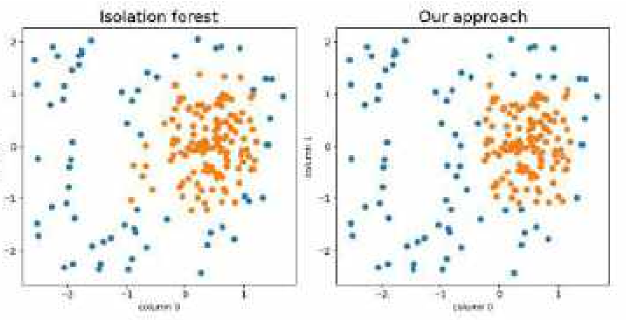


Fig. 3. Anomaly detection in comparison with *IsolationForest.*

IV. AnomalyTree for Data Analysis

AnomalyTree algorithm can be also applied as a preliminary step in data analysis. For instance, clustering algorithms will produce more accurate results if applied to a dataset with outliers already removed. No doubt, the most popular clustering algorithm is *k-means* clustering [8]. It is well-known for its simplicity, scalability, speed, interpretability and versatility. However, K-means is highly sensitive to outliers that can significantly affect the position and size of the clusters. To overcome this limitation, one can employ density-based algorithms like *DBSCAN* which are dedicated to cut off sparse regions not including them into clusters and treating them like outliers. However, DBSCAN is sensitive to parameter settings, as it requires the specification of two key parameters: epsilon (ε), representing the maximum distance between points in the same neighborhood, and the minimum number of points required to form a dense region (minPts). The algorithm's performance can be sensitive to these parameter values, and selecting appropriate values can be challenging. Besides, DBSCAN has difficulties with clustering high-dimensional data, its effectiveness decreases as the dimensionality of the data increases. In high-dimensional spaces, the distance between points tends to become less meaningful, making it harder for DBSCAN to capture clusters accurately. Apart from that, DBSCAN might be influenced by noise and outliers. Although DBSCAN is designed to handle noise and outliers, it can be sensitive to their presence. Noisy data or outliers can disrupt the density requirements and lead to the improper identification of clusters. And lastly, DBSCAN's time complexity is relatively high, especially for large datasets. The algorithm's performance depends on the number of data points and the neighborhood size, making it computationally expensive in scenarios with a vast amount of data.

A combination of simple and predictable *k-means* together with AnomalyTree density-based algorithm can become a solution to retrieve the best results from your clustering.

A. Synthetic data experiments

Let's test the AnomalyTree method in combination with K-means on a synthetic dataset with uniformly generated noise.

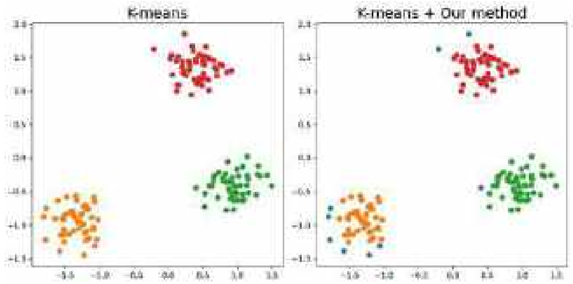


Fig. 4. Synthetic data without noise.

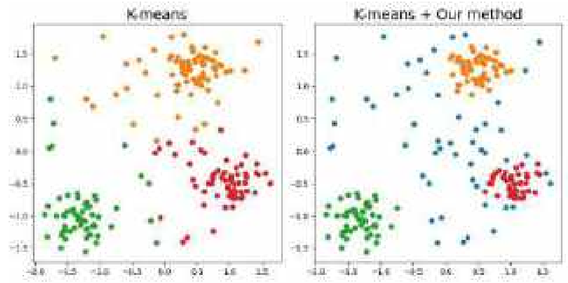


Fig. 5. Synthetic data with noise.

Figures 4 and 5 show how the AnomalyTree approach improves *k-means* clustering, efficiently cutting distant points and leaving pure clusters. In this way it enriches the amount of useful information gained from clustering.

In terms of *Silhouette score* [11], which is widely used to evaluate clustering performance on data with no labels, our approach was able to improve the results of clustering by about 20% – from 0.65 up to 0.84.

B. Real data experiments

Let's see an example of different approaches working on the famous 'Iris flower dataset' [9]. Suppose we want to find out what each of 3 types of flowers have in common and for more precise analysis we need to cut outliers, leaving only those flowers tightly placed near each other. For the sake of visualization, we fit all models to the data which was dimensionally reduced with Principal Component Analysis (*PCA)* [10], in our case from 4 to 2 dimensions.

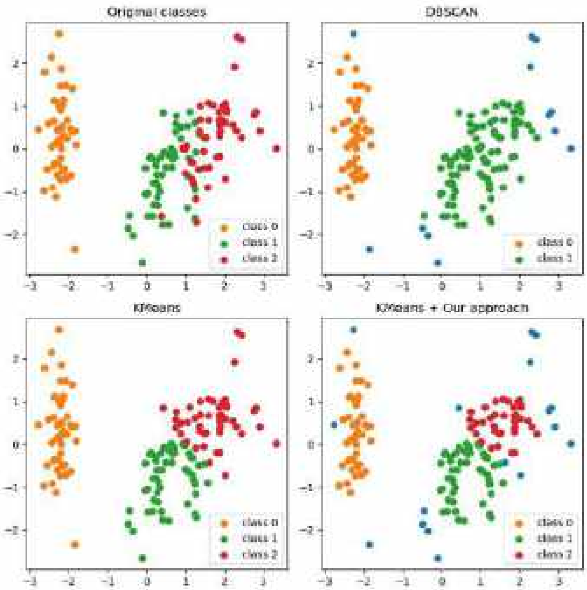


Fig. 6. Comparison of different clustering approaches.

As we see on Figure 6, *DBSCAN* was capable of cutting outliers, but unlike *k-means*, it failed to recognize all three initial clusters.

Let's take a look at popular clustering metrics to confirm or deny our visual assessment. We use *Silhouette Score* [11] and *Calinski-Harabasz Index* [12] for performance evaluation.

TABLE I. CLUSTERING SCORES COMPARISON

Performance evaluation			
	K-means	DBSCAN	**K-means + Our approach**
Silhouett e Score	0.51	0.52	**0.69**
Calinski-Harabasz Index	294	122	**441**

Table 1 proves AnomalyTree method to be effective in improving clustering results on real data as well. Besides, AnomalyTree in combination with *k-means* was able to cope with outliers no worse than *DBSCAN* and improved performance of *k-means* clustering.

Furthermore, it is also able to return insights on why we treat thrown away points as outliers explaining features and values of the cuts as well as their impact.

CONCLUSION

The goal of the paper is to present a density-based algorithm for outlier detection, called AnomalyTree. We provide open access to its implementation. The main idea behind the proposed approach is to partition the data space into both data-filled and empty regions at different levels of granularity to identify outliers.

To adapt the decision tree algorithm for outlier identification in an unsupervised manner, we leverage the idea of introducing non-existing points into the data space. Additionally, a modification to the purity function is made that anticipates the optimal partition.

To evaluate the effectiveness of AnomalyTree, extensive experiments were conducted. The results of these experiments demonstrated the efficacy of the proposed technique in achieving outlier detection objectives.

REFERENCES

[1] E. Martin, K. Hans-Peter, S. Jörg, X. Xiaowei, "A density-based algorithm for discovering clusters in large spatial databases with noise," Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press: Simoudis, Evangelos. Han, Jiawei. Fayyad, Usama M, 1996.

[2] M. Ankerst, M. Breunig, H.-P. Kriegel, J. Sander, "OPTICS: Ordering Points To Identify the Clustering Structure," ACM SIGMOD international conference on Management of data: ACM Press. pp. 49–60, 1999.

[3] E. Fix, J. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," Report No.4, USAF School of Aviation Medicine, Randolph Field, Tex., 19pp. 1951.

[4] M. Breunig, H.-P. Kriegel, R. Ng, J. Sander, "LOF: Identifying Density-based Local Outliers," Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data: Association for Computing Machinery, pp. 93–104. 2000.

[5] C. Cortes, V. Vapnik, “Support-vector networks,” Machine Learning 20(3), pp. 273-297, 1995.

[6] F. Liu, K. Ting, and Z. Zhou, “Isolation forest,” 2008 Eighth IEEE International Conference on Data Mining: IEEE, page 413-422, 2008

[7] B. Liu, Y. Xia and P. Yu, “Clustering Via Decision Tree Construction,” CIKM 2000, pp. 20-29, 2000.

[8] S. Lloyd, "Least square quantization in PCM," Bell Telephone Laboratories Paper, 1982.

[9] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, 7, pp. 179-188,1936.

[10] R. Vidal, Y. Ma, S. Sastry, “Generalized principal component analysis (GPCA),” IEEE Trans Pattern Anal Mach Intell. 27(12), pp. 1945–1959, 2005.

[11] P. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis," Comput. Appl. Math. 20, pp. 53-65, 1987.

[12] T. Calinski and J. Harabasz, “A dendrite method for cluster analysis,” Communications in Statistics. Theory and Methods, 3, pp. 1-27, 1987.

[13] H. Cheng, D. Xu, Sh. Yuan, X. Wu, “Fine-grained Anomaly Detection in Sequential Data via Counterfactual Explanations,” arXiv:2210.0414, 2022.

[14] Zh. Li, Y. Zhu, M. Leeuwen, “A Survey on Explainable Anomaly Detection,” arXiv:2210.06959, 2022.

[15] M. Jiang, Ch. Hou, A. Zheng, X. Hu, S. Han, H. Huang, X. He, P. Yu, Y. Zhao, “Weakly Supervised Anomaly Detection: A Survey,” arxiv:2302.04549, 2023.

Continuous Time Neural Network

Ihor Kutsevol
Postgraduate student,
Ivan Franko National University of Lviv
Lviv, Ukraine
ihor.kutsevol@lnu.edu.ua

Oleh Buhrii
Dr. of Sc. in Phys. and Math., Professor,
Ivan Franko National University of Lviv
Lviv, Ukraine
oleh.buhrii@lnu.edu.ua

***Abstract* — The architecture and learning process of a multilayer perceptron are considered. Presented a transformation of Res-Net into CTNN. The problem of silhouette recognition from white noise by these networks is defined and implemented using the Python programming language.**

Keywords* — *Artificial neural network (ANN), Multi-layer Perceptron, Continuous Time Neural Network (CTNN), Residual Neural Network (Res- Net), Ordinary Differential Equations Neural Network (ODE-Net).

I. Introduction

An artificial neural network is a model that tries to imitate the work of the human brain which consists of connected artificial neurons layers. Each neuron receives input data, processes it and transfers the result to the next neurons. Artificial neural networks (ANNs) are widely used. They are used to solve classification, regression problems, problems of objects recognition in images, natural language processing, etc. With the growth of computing capabilities of modern computers, the problems in which ANNs are used become more. This requires a revision of the classical ANN architectures and a theoretical justification of training algorithms. We will consider an illustrative example of ANNs constructions, will present how a discrete mathematical model of a multilayer perceptron is transformed into a neural network with continuous time, the work of which is modeled by a system of ordinary differential equations (ODEs). Usage of differential equations (ordinary and with partial derivatives) is a modern trend of ANN theory [1, 2, 3, 4] because of their practical value [5, 6]. The software implementation of the results was done, for example, using Julia programming language [7, 8]. We will look at some features of CTNN and implement the results in Python.

II. Materials and Methods

Multi-layer perceptron. Let's start our consideration with a certain modification of the results of [7]. In this article, we will consider a multi-layer perceptron schematically shown in Fig. 1. Let's explain the architecture of this network. It consists of an *input layer*

$$Input = I = \begin{pmatrix} i_1 \\ i_2 \end{pmatrix}, \tag{1}$$

one, so-called, *hidden layer*

$$Hidden = H = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \tag{2}$$

and an *output layer*

$$Output = O = \begin{pmatrix} o_1 \\ o_2 \end{pmatrix}, \tag{3}$$

Connections between layers are made with two sets of *weights*

$$Weight_1 = W = \begin{pmatrix} w_1 & w_2 \\ w_3 & w_4 \end{pmatrix}, \tag{4}$$

$$Weight_2 = V = \begin{pmatrix} v_1 & v_2 \\ v_3 & v_4 \end{pmatrix} = \begin{pmatrix} w_5 & w_6 \\ w_7 & w_8 \end{pmatrix}, \tag{5}$$

two sets of *biases*

$$Bias_1 = B = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, Bias_2 = \mathrm{B} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \tag{6}$$

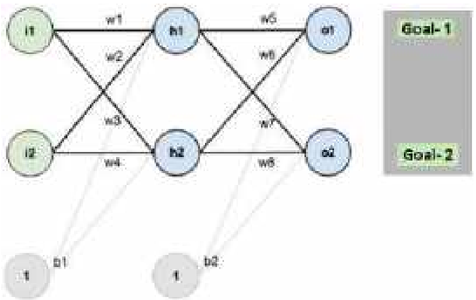


Fig. 1. Multi-layer perceptron

and, used twice, *activation function*

$$Sigmoid = s(x) = \frac{1}{1 + e^{-x}}, x \in \mathbb{R}. \tag{7}$$

Let us note that the sigmoid is a solution to the Cauchy problem of first-order ordinary differential equation

$$s' = s(1 - s), \tag{8}$$

$$s(0) = \frac{1}{2}. \tag{9}$$

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

Let's agree for convenience to use this function with a vector argument, assuming that

$$s\left(\begin{pmatrix}x_1\\x_2\end{pmatrix}\right)=\begin{pmatrix}s(x_1)\\s(x_2)\end{pmatrix},\begin{pmatrix}x_1\\x_2\end{pmatrix}\in\mathbb{R}^2. \tag{10}$$

The goal of perceptron training should be the implementation of “single point” set of input data

$$\left\{\begin{pmatrix}i_1\\i_2\end{pmatrix},Goal:=\begin{pmatrix}g_1\\g_2\end{pmatrix}\right\}. \tag{11}$$

The difference between *Output* and *Goal* is measured using the *loss function*

$$L\left(\begin{pmatrix}y_1\\y_2\end{pmatrix}\right)=\frac{1}{2}y_1^2+\frac{1}{2}y_2^2,\begin{pmatrix}y_1\\y_2\end{pmatrix}\in\mathbb{R}^2, \tag{12}$$

where

$$\begin{pmatrix}y_1\\y_2\end{pmatrix}:=\begin{pmatrix}g_1-o_1\\g_2-o_2\end{pmatrix}.$$

With introduced notations, the model is described as follows. At the *Hidden* level, an output signal should be

$$H=s(W\cdot I+B). \tag{13}$$

Similarly, an output signal at the *Output* level should be

$$O=s(V\cdot H+\mathrm{B}). \tag{14}$$

The problem of training the network is to find such weights and biases, which are, in a certain sense, the solution to the optimization problem

$$L\to min. \tag{15}$$

Function L has 12 variables: 4 weights in each of the matrix's W and V, *2* biases in each of the vector's B and B. The minimization takes place in the space $\mathbb{R}^{12}$. At the zero step, we select the initial values of the specified parameters. To obtain the next elements of the minimization sequence for the problem (15), we will use the *coordinate gradient descent* method. To find the corresponding new weights and biases we use the following formulas (for convenience, we show only four out of twelve formulas)

$$v_1^+=v_1-\eta\frac{\partial L}{\partial v_1}, \tag{16}$$

$$w_1^+=w_1-\eta\frac{\partial L}{\partial w_1}, \tag{17}$$

$$b_1^+=b_1-\eta\frac{\partial L}{\partial b_1}, \tag{18}$$

$$\beta_1^+=\beta_1-\eta\frac{\partial L}{\partial \beta_1}. \tag{19}$$

where η - some small parameter. As a result, according to the known rather small value of the parameter $\eta>0$, using the formulas of type (16) - (19), for given values of weights *W, V* and biases *B,* B, we construct new values W^+,V^+,B^+,B^+ such that

$$L(W^+,V^+,B^+,\mathrm{B}^+)\le L(W,V,B,\mathrm{B}).$$

Using the presented algorithm in a cycle way, in a certain number of steps, we will obtain such weights and biases that realize the permissible discrepancy between *Output* and *Goal*.

Continuous Time Neural Networks. We have considered the architecture of a perceptron with one hidden layer. Omitting the details, we can write it in the following form

$$Input\ \Rrightarrow Hidden \Rrightarrow Output,$$

or abbreviated

$$I\Rrightarrow H_1\Rrightarrow O.$$

To unify the notations, we will accept

$$H_0:=I,H_2:=O.$$

Then the formula (13) will take the form

$$H_1=s(W\cdot H_0+B), \tag{20}$$

and the formula (14)

$$H_2=s(V\cdot H_1+\mathrm{B}). \tag{21}$$

Since W and V are weights (albeit at different levels), it is natural to denote

$$W_1:=W,W_2:=V.$$

Similarly, let's redefine biases as

$$B_1:=B,B_2:=\mathrm{B}.$$

Then the formulas (20) - (21) can be combined into one:

$$H_{t+1}=s(W_{t+1}\cdot H_t+B_{t+1}),t=0,1. \tag{22}$$

A neural network (22) consists of the so-called *plain block*, in which previous information is considered only in the argument of the activation function. It appears that to speed up the learning process (which is especially relevant for complex tasks, for example, image recognition), instead of a plain block (22), it is more appropriate to consider a so-called residual block

$$H_{t+1}=H_t+s(W_{t+1}\cdot H_t+B_{t+1}),t=0,1. \tag{23}$$

Such a network is called *Residual Neural Network (RNN)*. Let's consider the scheme of problem solving, in which the connections between levels (22) are replaced by (23). Of course, the specific view of the formulas for training the network will change. However, the algorithm will remain the same. It will consist of finding eight weights and four biases to "correctly" recognize the two-dimensional input vector.

When the input information becomes more complicated (for example, the dimension of the H_0 vector increases, there are more *Input* - *Goal* pairs, etc.), the learning process will become more and more cumbersome and will be poorly implemented even on powerful computers. One of the options for solving this problem and improving the quality of the network is to increase the number of hidden layers without changing their structure. Indeed, if instead of (23) we consider the network

$$H_0 \Rrightarrow H_1 \Rrightarrow \cdots \Rrightarrow H_t \Rrightarrow H_{t+1} \Rrightarrow \cdots \Rrightarrow H_N,$$

which has *N - 1 Hidden level*, moreover,

$$H_{t+1} = H_t + f(H_t, W_{t+1}, B_{t+1}), t = \overline{0, N-1}, \qquad (24)$$

where

$$f(H_t, W_{t+1}, B_{t+1}) \coloneqq s(W_{t+1} \cdot H_t + B_{t+1}), \qquad (25)$$

then the number of *"unknown parameters" = "weights and biases"* would increase, but the procedure for finding them would remain standard. Let's combine the weights W_{t+1} and the bias B_{t+1} into one notation *u(t)* (this is a vector whose elements are the elements of the matrix W_{t+1} and the vector B_{t+1} and introduce the following additional notations:

$$\Delta t = 1 = (t+1) - t, x(t) = H_t,$$

$$g\big(x(t), u(t)\big) = f(H_t, W_{t+1}, B_{t+1}).$$

Then, instead of (24), we can write the network architecture in the form

$$x(t+1) = x(t) + \ g\big(x(t), u(t)\big)\Delta t, t = \overline{0, N-1}. \ (26)$$

The formula (26) is an implementation of Euler's method of constructing a solution of the Cauchy problem for a system of ordinary differential equations

$$x' = g(x, u), \qquad (27)$$
$$x(0) = x_0, \qquad (28)$$

where the vector x_0 depends on the data H_0. Since the solution of the problem (27) - (28) under certain conditions satisfies Volterra's integral equation of the 2nd order

$$x(t) = x(0) + \int_0^t g(x(\tau), u(\tau))d\tau, t \in [0, N], \qquad (29)$$

then the procedure for constructing *N* levels of the network can be replaced by one formula of the type (29):

$$x(N) = x(0) + \int_0^N g(x(\tau), u(\tau))d\tau. \qquad (30)$$

The process of network training will consist in the minimization of some objective function

$$L\big(x(N)\big) = L\left(x(0) + \int_0^N g(x(\tau), u(\tau))d\tau\right) =$$
$$= L\big(ODESolver(x(0), g, 0, N, u)\big). \qquad (31)$$

where the expression *ODESolver* formally means the application of the Cauchy solution search operator for a system of ordinary differential equations in some programming languages (for example, in Python). The main technical problem in training such a network is trying to differentiate by parameter (this is called backpropagation) "inside" the *ODESolver* operator. However, this problem can already be solved technically. Thus, we need to solve the following problem:

$$J(x, u) \coloneqq L\big(x(N)\big) \to min, \qquad (32)$$
$$x' = g\big(x(t), u(t)\big), t \in [0, N], \qquad (33)$$
$$x(0) = x_0, \qquad (34)$$
$$u(t) \in U, t \in [0, N]. \qquad (35)$$

This is one of the classic problems of optimal control theory, which is called the *terminal control problem*. The function (32) is called the *Mayer functional*, the vector function *x* - the *phase coordinates* or *allowable trajectories* of the problem, the vector function *u* - *permissible control* tasks. Thus, *x* is the solution to the Cauchy problem (33) - (34). The values of *u* belong to a certain set of admissible controls *U*.

III. Software implementation of algorithms

Consider the problem of silhouette recognition from white noise using the residual neural network Res-Net (23) and the above-mentioned ODE-Net model (32) - (35). In other words, we need to find such a set of weights that would transform the white noise into a silhouette, would make the model see the silhouette where it is not. Let there be a "one-point" set of input data

$$\left\{I_1 = I = \begin{pmatrix} i_1 \\ \cdots \\ i_{1024} \end{pmatrix}, G_1 = G = \begin{pmatrix} g_1 \\ \cdots \\ g_{1024} \end{pmatrix}\right\}, \qquad (36)$$

namely, 32 × 32 sized images, which are transformed into a vector with *1024* coordinates. Here, *I* is a random vector that is white noise, and *G* is a certain silhouette (see Fig. 2).

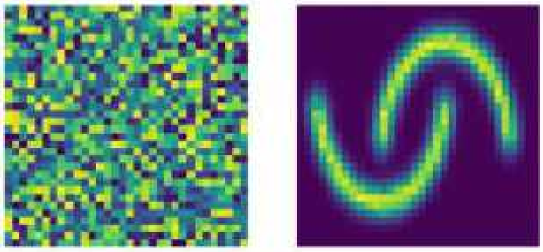


Fig.2. *I* is the input noise image, G is the silhouette

To do this, we implement the residual neural network (Res-Net) and neural network with ordinary differential equations (ODE-Net) using the Python programming language and a machine learning framework for the transformation of numerical functions *Google JAX*. This platform provides the possibility of applying the Cauchy problem solution search operator for the system of ordinary differential equations (*ODESolver* in (31)), compiling the source code into machine code, finding the gradient, etc. Each model is built in several steps.

Step 1. Define the hyperparameters for model

$$depth_{resnet} = 5; depth_{odenet} = 3;$$

$$layers_{resnet} = [32 * 32] * depth_{resnet};$$

$$layers_{odenet} = [32 * 32 + 1] * (depth_{odenet} - 1) + \\ + [32 * 32];$$

$$scale_{resnet} = scale_{odenet} = 0.001;$$

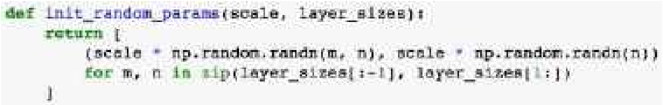
$$alpha_{resnet} = alpha_{odenet} = 0.01;$$

$$iters_{resnet} = 1000; iters_{odenet} = 100;$$

where *depth* - is the total number of layers, *layers* - a list that contains the number of neurons in the corresponding layer, *scale* - the number by which the generated random weights and biases are multiplied, *alpha* - learning rate coefficient, *iters* - the number of training iterations.

Step 2. Generate random values of weights and biases of the required dimension, considering the multiplier *scale*.

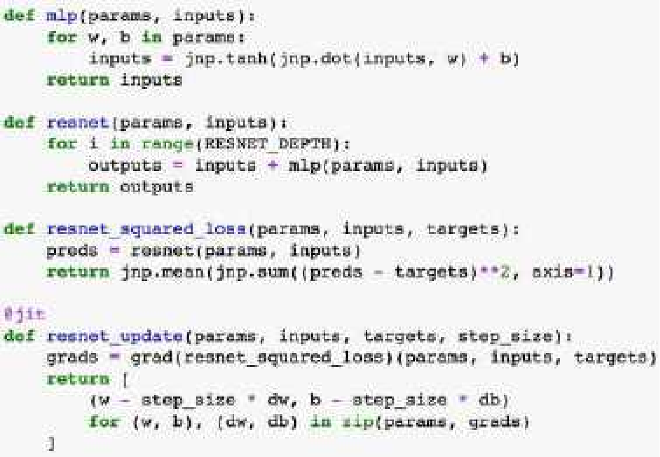
The function shown in Fig. 3 is responsible for this.



```
def init_random_params(scale, layer_sizes):
    return [
        (scale * np.random.randn(m, n), scale * np.random.randn(n))
        for m, n in zip(layer_sizes[:-1], layer_sizes[1:])
    ]
```

Fig. 3. Function to initialize random weights and biases

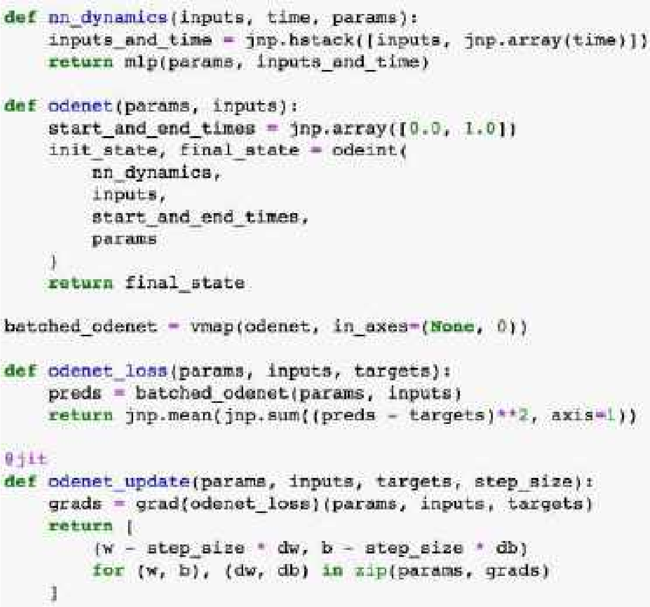
Step 3. Define the architectures of Res-Net and ODE-Net models and the weights adjustment method as follows.



```
def mlp(params, inputs):
    for w, b in params:
        inputs = jnp.tanh(jnp.dot(inputs, w) + b)
    return inputs

def resnet(params, inputs):
    for i in range(RESNET_DEPTH):
        outputs = inputs + mlp(params, inputs)
    return outputs

def resnet_squared_loss(params, inputs, targets):
    preds = resnet(params, inputs)
    return jnp.mean(jnp.sum((preds - targets)**2, axis=1))

@jit
def resnet_update(params, inputs, targets, step_size):
    grads = grad(resnet_squared_loss)(params, inputs, targets)
    return [
        (w - step_size * dw, b - step_size * db)
        for (w, b), (dw, db) in zip(params, grads)
    ]
```

Fig. 4. The Res-Net model in Python programming language

In Fig. 4 the *mlp* function implements a nonlinear transformation of input signals using the activation function (in this case *tanh*), the *resnet* function actually, describes the residual neural network, *loss* - the target function (12) in $\mathbb{R}^{32\times32}$, *update* - defines a gradient descent method to adjust weights and biases with the specified hyperparameter *alpha*.

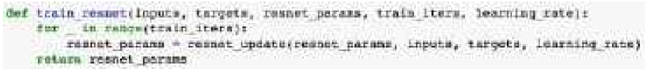


```
def nn_dynamics(inputs, time, params):
    inputs_and_time = jnp.hstack([inputs, jnp.array(time)])
    return mlp(params, inputs_and_time)

def odenet(params, inputs):
    start_and_end_times = jnp.array([0.0, 1.0])
    init_state, final_state = odeint(
        nn_dynamics,
        inputs,
        start_and_end_times,
        params
    )
    return final_state

batched_odenet = vmap(odenet, in_axes=(None, 0))

def odenet_loss(params, inputs, targets):
    preds = batched_odenet(params, inputs)
    return jnp.mean(jnp.sum((preds - targets)**2, axis=1))

@jit
def odenet_update(params, inputs, targets, step_size):
    grads = grad(odenet_loss)(params, inputs, targets)
    return [
        (w - step_size * dw, b - step_size * db)
        for (w, b), (dw, db) in zip(params, grads)
    ]
```

Fig. 5. The ODE-Net model in Python programming language

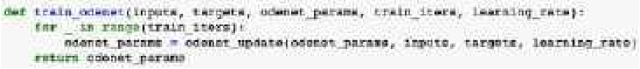
In Fig. 5 the *dynamics* function also implements a nonlinear transformation of the input signals using an activation function, but also considering the time interval, while the *odenet* function describes a continuous-time neural network.

Step 4. Determine the functions of iterative learning, during which we adjust the weights and biases of the models using the gradient descent method.



```
def train_resnet(inputs, targets, resnet_params, train_iters, learning_rate):
    for _ in range(train_iters):
        resnet_params = resnet_update(resnet_params, inputs, targets, learning_rate)
    return resnet_params
```

Fig. 6. Description of Res-Net model training using Python



```
def train_odenet(inputs, targets, odenet_params, train_iters, learning_rate):
    for _ in range(train_iters):
        odenet_params = odenet_update(odenet_params, inputs, targets, learning_rate)
    return odenet_params
```

Fig. 7. Description of ODE-Net model training using Python

We implement training using the *for* loop, update the weights and biases *iters* times and return the latest updated values (Fig. 6-7). We call the functions and get the trained models.

IV. Visualization of results

After obtaining the weights and biases for each model after the last iteration of training, we compare and visually evaluate the training results.

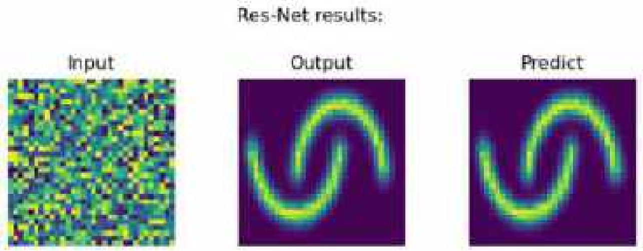


Fig. 8. Input, Expected and Predicted by Res-Net Images

It can be noted in Fig. 8 that the residual neural network perfectly coped with the task of silhouette recognition on the *Input* image. The value of the objective function for this model is *0.001*, which indicates that each corresponding pixel in the *Output* and *Predict* images has almost the same saturation.

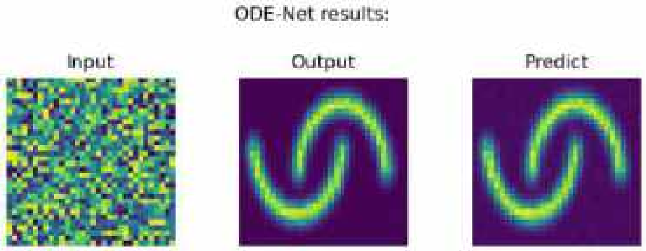


Fig. 9. Input, Expected and Predicted by ODE-Net Images

The continuous time neural network also coped with this task, but it is easy to notice in Fig. 9 that in some places the *Output* and *Predict* images do not exactly match. This is evidenced by the value of the objective function, which is equal to *0.05*. However, it is worth noting that the ODE-Net model was trained for 100 iterations, which is 10 times less than Res-Net. Training took place at the same values of *alpha = 0.01* - the learning rate coefficient. Another important point is the training time of the models. The Res-Net architecture clearly outperforms ODE-Net on this metric, as the residual network took only 7 seconds to train, while ODE-Net took 3 minutes. Obviously, this is caused by the large number of calculations, which mainly fall on the function *odeint* (*ODESolver* in (31)), which needs improvements. In addition, solving the corresponding problem of optimal control can traditionally be reduced to the use of the Pontryagin maximum principle, which should speed up the learning process. This will make it possible to widely use differential equations in the modeling of complex processes without spending a lot of resources on it.

Conclusion

The architecture and learning method of the multilayer perceptron was demonstrated (Fig. 1). After that, the transition from residual to continuous-time neural network was shown (26). Considered the problem of silhouette recognition from white noise using Res-Net (23) and ODE-Net (32) - (35) models. The obtained results (Fig. 8 - 9) indicate the competitiveness of two models, but there are advantages and disadvantages of using each of them. The Res-Net model proved itself to be excellent. Disadvantages include the extremely large number of iterations required for training. However, the iterations were done quickly, which is an advantage over ODE-Net, which in turn takes much more time for a smaller number of iterations. This is caused by *odeint* function calls, which consume a significant amount of computing resources.

References

[1] R.T.Q. Chen, et al., "Neural Ordinary Differential Equations", Advances in neural information processing system 31, 2018.

[2] Ch. Rackauckas, et al., "Universal differential equations for scientific machine learning", arXiv preprint arXiv:2001.04385, 2020.

[3] L. Ruthotto Eldad H., "Deep neural networks motivated by partial differential equations", Journal of Mathematical Imaging and Vision 62, 2020, pp. 352-364.

[4] R. Hasani, et al., "Closed-form continuous-time neural networks", Nature Machine Intelligence, 2022, pp. 1-12.

[5] M. Nunez, et al., "Forecasting virus outbreaks with social media data via neural ordinary differential equations", MedRxiv, 2021.

[6] C. Vorbach, et al., "Causal navigation by continuous-tiime neural networks", Advanced in Neural Information Processing Systems 34, 2021, pp. 12425-12440.

[7] M.M. Saeed, "Ordinary Differential Equations Neural Networks: Mathematics and Application using Diffeqflux.jl", https://scholarworks.arcadia.edu/senior_theses/43/, 2019, pp. 1 – 20.

[8] C. Rackauckas, et al., "Diffeqflux.jl - A julia library for neural differential equations", arXiv preprint arXiv:1902.02376, 2019.

Modification of Combined Unsupervised-Supervised Cascade Scheme for Small Biomedical Data Classification

Ivan Izonin, *IEEE Senior Member*
Department of Artificial Intelligence
Lviv Polytechnic National University
Lviv, Ukraine
0000-0002-9761-0096

Roman Tkachenko
Department of Publishing Information Technologies
Lviv Polytechnic National University
Lviv, Ukraine
0000-0002-9802-6799

Marian Sydor
Department of Artificial Intelligence
Lviv Polytechnic National University
Lviv, Ukraine
0000-0002-7034-8075

Stephane Chretien,
ERIC Laboratory and UFR ASSP,
Université Lumiere Lyon 2
Bron, France,
0000-0002-4544-1315

Iryna Pliss
Control Systems Research Laboratory
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
0000-0001-7918-7362

Taras Vovk
Department of Artificial Intelligence
Lviv Polytechnic National University
Lviv, Ukraine
0009-0003-8418-036X

***Abstract* — The task of improving classification accuracy in the case of analysis of short datasets is an important problem, in particular, for express diagnostics in biomedicine. Hybrid methods, including ensemble methods, are gaining more and more popularity for their solution. This paper proposes a modification of the unsupervised-supervised cascade scheme, the purpose of which is to reduce the dimensionality of the extended dataset intended for analysis. The proposed modification includes the use of the clustering method for data pre-processing, as well as the parallel application of a linear classifier at the first level of the cascade. As a result of this approach, the initial inputs of the task are replaced by only one attribute, the output signal of the linear classifier, which, together with the markers belonging to each data cluster, forms a new dataset for training the final classifier on the second cascade level. The simulation was performed on a real dataset using six different linear classifiers. Through comparison, it was established both a reduction in the time resources required for training the method and an increase in the accuracy during the analysis of small sets of biomedical data.**

***Keywords* — express diagnostics, ensemble learning, unsupervised-supervised learning, small data approach, linear methods, cascade scheme**

I. Introduction

Today, biomedical engineering includes more than 18 areas of research [1]. The modern development of artificial intelligence provides the possibility of applying its latest achievements to solve various tasks in each of these areas. One of these tasks is data classification.

Features of biomedical datasets are the presence of a large number of features in a given dataset [2]. This is explained by the need to take into account both medical and biological and engineering and technical characteristics that affect the dependent characteristic [3]. In addition, today many tasks of biomedical engineering are characterized by a limited amount of available data [4]. This reduces the efficiency or makes the intellectual analysis of such data impossible. A big problem is a short data set with a large number of attributes in each vector from the set [5], [6]. If we take into account the condition that the number of features should be at least 100 times less than the number of observations [7], it imposes many problems when solving various problems of intellectual analysis based on such biomedical datasets.

One of the possible options for increasing the classification accuracy, in particular by linear methods, is to increase the dimensionality of the input data space of the task, in particular with the use of clustering [8]. However, in the case of the analysis of a short set of biomedical data, taking into account its already typical high dimensionality, the use of such methods is accompanied by several limitations.

However, taking into account the fact that similar methods demonstrate very high accuracy [8], this paper aims to modify the unsupervised-supervised cascade scheme to reduce the dimensionality of the extended dataset intended for analysis.

Therefore, the main contribution of this paper can be summarized as follows:

- The two-step data approximation method [8] was adapted for the case of solving a classification task, which, due to the expansion of the input data space of the task with markers of belonging to each cluster, determined by an additional clustering procedure, ensures an increase in the accuracy of the work of linear classifiers.
- The two-step method of data classification has been modified, which, unlike the existing one, replaces all initial independent attributes with the predicted value of the sought variable due to the additional use of a classifier in the first step of the method. This approach reduces the training time of the final classifier while maintaining or even increasing the accuracy of its work.

II. State-of-the-Arts

The idea of using clustering and classification together is not new. There are many articles and applied problems where

The EURIZON Fellowship Program "Remote Research Grants for Ukrainian Researchers" funded this research under grand № 138.

the combination of these two approaches provides a significant increase in accuracy or performance.

In particular, in [9] the authors used unsupervised learning for pre-processing data in bioinformatics. The fact is that this area is characterized by a large amount of data that is difficult to label. Some errors and inaccuracies affect the reliability of the data for the classifier [10]. That is why the author developed an approach to the formation of a more reliable dataset for the classifier using unsupervised learning.

In [11] an equally interesting task was solved. The authors developed a new approach to feature selection using unsupervised learning. It is based on the idea of a random forest, but the Unsupervised Feature Selection mechanism is used here. Modeling of the proposed approach was carried out using 19 datasets. The results demonstrate a significant increase in the accuracy of the method in comparison with analogs.

In [12], [13] methods of increasing data classification accuracy using clustering have been developed. The first step of the method involves using clustering to select compact sets of points. In the next step, a classifier works in each such set. Due to this, the overall classification accuracy on the entire dataset increases significantly. However, in cases of the existence of a small number of observations in one of the selected clusters, the application of a classifier based on machine learning may be significantly limited.

To overcome the problem described above, in [8] another scheme of combining clustering and, in this case, regression is considered. In particular, pre-processing of data that takes place using clustering serves as additional information that will be included in the general dataset. In particular, each vector will be expanded with markers belonging to each of the identified clusters. This approach solves the problem of the previous method, significantly increasing the prediction accuracy. However, it also increases the computational complexity of the classifier and can provoke overfitting. In addition, it is not always suitable for the analysis of short datasets.

III. Modification of Combined Unsupervised-Supervised Cascade Scheme

This section describes the adapted method (basic is [8]) and the two-step method modified in this paper for solving the data classification task. In addition, the section provides a brief description of the linear classification methods that were used for modeling to determine the effectiveness of the proposed approach. In addition, methods for determining the optimal number of clusters for the data clustering procedure required by existing and modified methods are described here.

A. *Adaptation of the existing method for solving classification tasks*

The authors in [8] developed a method for the combined use of clustering and regressors based on machine learning algorithms to improve the accuracy of solving approximation tasks of tabular datasets. The method is based on the principle of expanding the space of input data due to the previous use of an additional procedure of clustering the input dataset. Clustering is performed on the training dataset using the *k-means* method. As a result of its execution, we get a set of

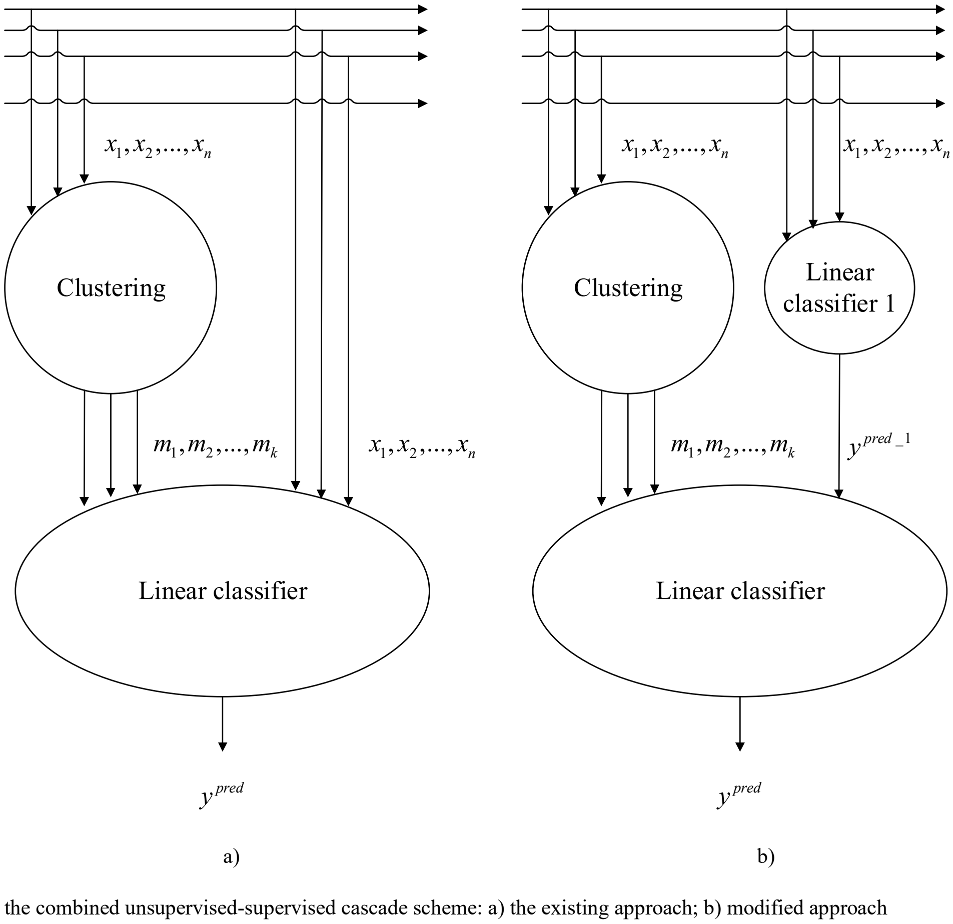


Fig. 1. Flowchart for the combined unsupervised-supervised cascade scheme: a) the existing approach; b) modified approach

clusters with each vector of the training data set belonging to one of them. An important condition for our adaptation (exactly for solving the classification task) is that the number of clusters must be greater than the number of previously known classes of the task.

The next step of the method is the formation of a new dataset by expanding each data vector with markers belonging to each of the specified clusters found in the previous step (1 – when the vector belongs to a cluster, 0 – in all other cases). In this way, we get a significantly expanded data set (larger number of attributes), which is used to train the final classifier.

In the application mode of the method, in the first step, we need to determine the belonging of each vector from the test dataset to the clusters defined at the training stage. This procedure takes place by determining the smallest Euclidean distance to each of the cluster centers determined at the training stage. Next, each vector of the test sample is also expanded with additional attributes, and markers belonging to each of the defined clusters (1 – when the vector belongs to a cluster, 0 – in all other cases). We apply the obtained expanded vectors to the previously trained final classifier. It produces the final result. Fig. 1 shows the structural diagram of this approach.

As can be seen from the description of the method and also from Fig. 1 a, the main disadvantage of this method is a significant expansion of the dimensionality of the input data space of the task. This significantly increases the duration of the training procedure of the final classifier and can worsen the generalization properties of the method. However, the main disadvantage of this method during the analysis of short datasets is a significant increase in the number of features. If we take into account the rule that the number of features should be at least 100 times less than the number of vectors in a current dataset [7], the above-described method imposes a series of limitations on the size of a short dataset, as well as the number of clusters that can be used by this method.

To eliminate all the above-mentioned shortcomings during the analysis of short datasets, this paper proposes a modification of the above-mentioned method.

B. Proposed modification

The modified method is also based on the idea of using clustering for pre-processing of a given data set. It also uses the k-means method to assign each observation to a defined cluster. Thus, as a result of clustering, we get a set of clusters, which should also be larger than the number of known classes of the problem. However, in the modified method, in parallel with clustering, a classification procedure using the first method of machine learning is introduced (Fig. 1.b).

As a result of the parallel execution of both of the above-described procedures, a new dataset is formed. In this case, all initial features of the task are replaced by one attribute - the output signal of the first classifier. Also, markers of belonging of each vector of the training sample to each of the obtained clusters are added to it. Next, the training procedure of the final classifier takes place on a significantly smaller dataset compared to the basic method described above.

In the application mode, we use the first classifier to obtain the output signal for each vector of the test sample and replace the initial features of the corresponding vector with the found output signal. To them are added markers of belonging of the current vector from the test sample to each of the known clusters. This happens based on determining the smallest Euclidean distance between the current vector and each of the known cluster centers. The obtained vectors are applied to the previously trained final classifier to obtain the observation's membership in one of the classes defined by the task.

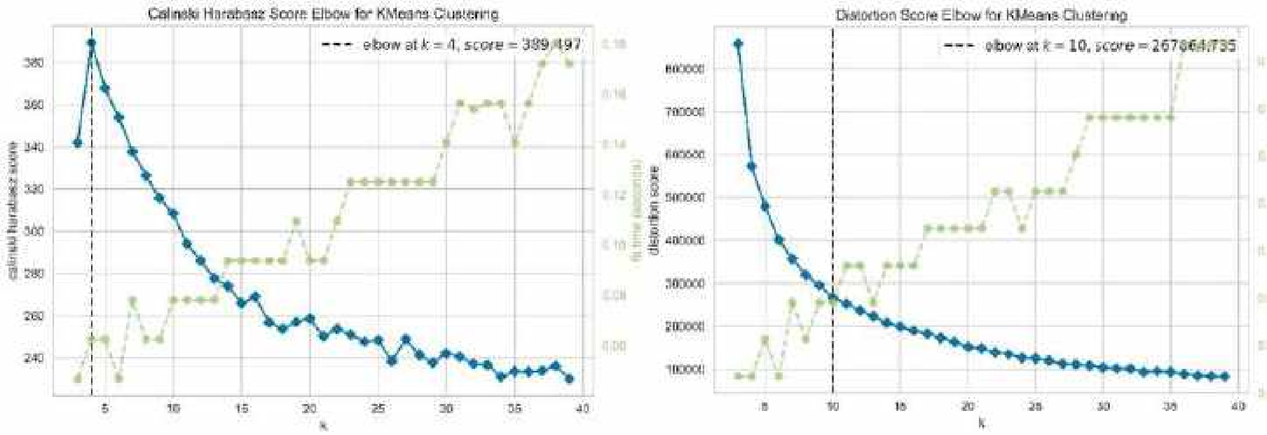
The structural diagram of this approach is presented in Fig. 1 b. An obvious advantage of such a step is an insignificant increase in the number of features of the extended dataset, which will provide the possibility of analyzing short datasets; will increase the performance of the developed two-step classifier; and even increase the accuracy of its work.

C. Basic ML-algorithms used for modeling

To test the effectiveness of the proposed approach, we used six different linear classifiers implemented in the scikit-learn (sklearn) library of the Python language [14]:

1) LogisticRegression is a simple algorithm for binary classification that creates a linear combination of input features and their weights and then passes that combination through a logistic function that transforms it into an interval between 0 and 1. This is interpreted as the probability that an object belongs to a certain class.

2) SVC (linear kernel) is used to separate two classes of objects by finding the optimal hyperplane in the feature



a)

b)

Fig. 2 Optimal k-mean's cluster number for the investigated dataset using: a) Calinski Harabasz method; b) Distortion method

space. The main idea is to find such a hyperplane that maximally separates objects of two classes.

3) LinearSVC – the method is identical to the previous one, but its implementation differs from (2) and provides a higher speed of operation.

4) RidgeClassifier - is a variation of the linear regression method that uses Ledge regularization for the classification task. The main goal is to find the optimal linear hyperplane that separates data from different classes.

5) GaussianNB is used for classification by applying a naive Bayesian approach with the assumption that the distribution of features in each class is normal (Gaussian). The main idea is to calculate the probabilities of the object belonging to different classes based on the probabilities of the appearance of features in each class.

6) SGDClassifier uses stochastic gradient descent to train a classification model. The main idea is to gradually update model parameters depending on the gradient of the loss function based on small data samples (mini-packets).

D. Methods for determining the optimal number of clusters

The basic and modified methods use a clustering procedure for pre-processing the data. It is performed using the *k-means* method, the main drawback of which is the need to manually select the number of clusters. To avoid this shortcoming, this paper uses two methods for determining the optimal number of clusters:

- Calinski-Harabasz is a method that consists in comparing the variance between clusters and the variance within clusters. A detailed description of the method can be found in [15].
- Distortion is a method that consists in evaluating how similar the objects within the clusters are to each other, as well as finding the "elbow point" on the graph that indicates a change in the intracluster variance variable. A detailed description of the method can be found in [16].

IV. Modeling and Results

This section describes the modeling procedure with the selection of the optimal parameters of the method and also presents the obtained results. The modeling was carried out on a computer with the following parameters: Processor: Intel(R) Core i7-8750H; Number of processor cores: 6: Operating frequency of the processor: 2.20GHz; RAM size: 32 GB; Operating system: Windows 10 Home; Python version: 3.9; Integrated development environment: PyCharm 2019.2.6 (Professional Edition). Total Accuracy indicator and training time were used to assess the accuracy of the studied methods.

A. Dataset descriptions

We used the dataset “Heart Attack Prediction” [17] for modeling the proposed approach. Here, the heart attack predicting task based on 14 independent attributes is solved. After preliminary processing, which consisted in removing columns containing many gaps, converting categorical features into numerical ones, as well as performing feature selection procedures, the final dataset contained 18 independent attributes and 294 observations.

B. Optimal k-mean’s cluster number

To determine the optimal number of clusters of the *k-means* clustering method, two methods were studied in the paper. Fig. 2 shows their results. It should be noted that the Calinski-Harabasz method demonstrates 4 clusters for a current dataset. At the same time, the use of the Distortion method does not provide an opportunity to determine the optimal number of clusters. Therefore, in the paper, a search was carried out on the interval [4, 10] with step 1. It was experimentally determined that 6 clusters provide the highest accuracy in solving the classification task by the method modified in this paper.

C. Investigating the effect of data normalization on the accuracy of all methods

The effectiveness of machine learning methods largely depends on the characteristics of the dataset [18]. In particular, some independent variables can acquire large values, while others - are very small. Accordingly, this will affect the accuracy of the selected classifier.

In this paper, we conducted experimental studies and compared the effectiveness of the developed and basic methods when applying six different linear machine learning methods with basic parameters. Also, we used three different normalization methods and a variant without data normalization.

Table 1 summarizes the results of this experiment.

TABLE I. Obtained Results for Different Normalization Techniques

ML model	Basic ML algorithm		Basic cascade method		Modified cascade	
	Training time (seconds)	*Accuracy*	*Training time (seconds)*	*Accuracy*	*Training time (seconds)*	*Accuracy*
Without normalization						
LogisticRegression	0,006883	0,8475	0,004220	0,8475	0,004191	0,8475
SVC (linear kernel)	0,006530	0,5424	0,006462	0,5424	0,010451	0,5763
LinearSVC	0,064928	0,7627	0,051804	0,8136	0,035782	0,7797
RidgeClassifier	0,144964	0,8305	0,043892	0,8475	0,055393	0,8305
GaussianNB	0,003866	0,8475	0,003547	0,8305	0,002397	0,8305
SGDClassifier	0,058517	0,5593	0,045070	0,5593	0,053689	0,5763
StandartScale						
LogisticRegression	0,003953	0,8644	0,003895	0,8644	0,002697	0,8644

ML model	**Basic ML algorithm**		**Basic cascade method**		**Modified cascade**	
	Training time (seconds)	***Accuracy***	***Training time (seconds)***	***Accuracy***	***Training time (seconds)***	***Accuracy***
SVC (linear kernel)	0,009777	0,8475	0,010116	0,8305	0,005557	0,8475
LinearSVC	0,066730	0,8644	0,056780	0,7966	0,035958	0,8644
RidgeClassifier	0,044398	0,8305	0,048889	0,8136	0,043394	0,8305
GaussianNB	0,002107	0,5593	0,002581	0,5593	0,002340	0,5593
SGDClassifier	0,047664	0,8305	0,042829	0,7966	0,149941	0,8305
MaxAbsScale						
LogisticRegression	0,002548	0,8305	0,003661	0,8305	0,002577	0,8305
SVC (linear kernel)	0,007695	0,8136	0,007601	0,8305	0,004236	0,8305
LinearSVC	0,045668	0,8305	0,064881	0,8305	0,053417	0,8305
RidgeClassifier	0,040368	0,8305	0,043315	0,8305	0,040007	0,8305
GaussianNB	0,002133	0,6949	0,002268	0,8305	0,002353	0,8305
SGDClassifier	0,038464	0,8305	0,045740	0,8305	0,039975	0,8305
MinMaxScale						
LogisticRegression	0,002682	0,8305	0,003143	0,9153	0,002681	0,9153
SVC (linear kernel)	0,007980	0,8136	0,008082	0,9153	0,004528	0,9153
LinearSVC	0,045368	0,8305	0,042081	0,8983	0,044613	**0,9153**
RidgeClassifier	0,044468	0,8305	0,051226	0,9153	0,045198	0,9153
GaussianNB	0,002482	0,6610	0,003986	0,8644	0,003406	**0,8814**
SGDClassifier	0,003164	0,8305	0,005432	0,8983	0,005191	**0,9153**

As can be seen from Table 1, MinMaxScaler demonstrated the highest accuracy among all other options. The results obtained using this scaler were used to compare the performance of all the methods studied in the paper.

V. Comparison and Discusion

The comparison was made using six well-known linear machine learning methods in the following modes:

- Using linear classifiers on the initial data set (training took place on a set of 18 initial independent attributes);
- Using linear classifiers within the framework of the basic method (training took place on a set of 24 independent attributes: 18 initial and 6 cluster labels);
- Using linear classifiers within the framework of the modified method (training took place on a set of 7 independent attributes: 1 output of the first linear classifier and 6 cluster labels).

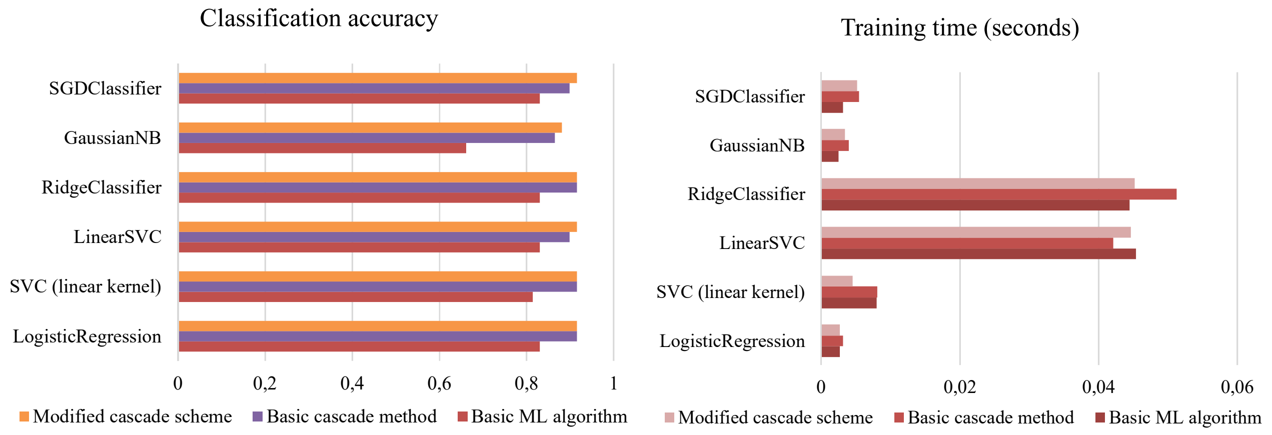
Fig. 3 a and 3 b summarize the accuracy and speed results of all six classifiers for the three proposed modes.

As can be seen from Fig. 3 linear machine learning methods show the lowest classification accuracy. The adapted basic method shows a significant increase in accuracy (from 9 to 20%). The modified method shows either the same accuracy as the basic one or an increase in accuracy by an average of 1.5% compared to the basic one. However, the main advantage of the modified method is to reduce the size of the data set for submission to the final classifier. This provides an increase in the generalization properties of the classifier, as well as a significant reduction in the duration of the training procedure in comparison with the basic method [8].

Conclusions

This paper adapts a two-step data classification method and presents its modification for intelligent analysis of short sets of biomedical data. The proposed modification includes the use of the clustering method (in particular, the *k-means* method) for data pre-processing, as well as the parallel application of an additional linear classifier. As a result of this approach, the initial inputs of the task are replaced by only one attribute, the output signal of the first linear classifier, which, together with the markers belonging to each cluster, forms a new dataset for training the final classifier.

The article solves the heart attack predicting task. The authors determined the optimal operating parameters of the modified method. The comparison is carried out in three modes: using linear classifiers on the initial dataset, within the framework of the basic method, and within the framework of the modified method. The results of the comparison show that the modified method shows a significant increase in accuracy (up to 22%) compared to basic machine learning algorithms. This approach also contributes to the reduction of the size of the dataset to be submitted to the final classifier, which leads to the acceleration of the work of the developed classifier. This



a)

b)

Fig. 3 Performance indicators for basic ML-algorithms, basic cascade method and moditied cascade scheme: a) classification accuracy; b) training time (seconds)

advantage will become quite noticeable during the analysis of significantly larger sets of biomedical data.

Acknowledgment

The British Academy's Researchers at Risk Fellowships Programme support this research.

References

[1] L. J. Street, Introduction to biomedical engineering technology, Fourth edition. Boca Raton: CRC Press, 2022.

[2] Y. Tolstyak and M. Havryliuk, 'An Assessment of the Transplant's Survival Level for Recipients after Kidney Transplantations using Cox Proportional-Hazards Model', CEUR-WS.org, vol. 3302, pp. 260–265, 2022.

[3] A. Altameem, V. Kovtun, M. Al-Ma'aitah, T. Altameem, F. H, and A. E. Youssef, 'Patient's data privacy protection in medical healthcare transmission services using back propagation learning', Computers and Electrical Engineering, vol. 102, p. 108087, Sep. 2022, doi: 10.1016/j.compeleceng.2022.108087.

[4] A. Salazar, L. Vergara, and G. Safont, 'Generative Adversarial Networks and Markov Random Fields for oversampling very small training sets', Expert Systems with Applications, vol. 163, p. 113819, Jan. 2021, doi: 10.1016/j.eswa.2020.113819.

[5] I. Krak, V. Kuznetsov, S. Kondratiuk, L. Azarova, O. Barmak, and P. Padiuk, 'Analysis of Deep Learning Methods in Adaptation to the Small Data Problem Solving', in Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making, S. Babichev and V. Lytvynenko, Eds., in Lecture Notes on Data Engineering and Communications Technologies, vol. 149. Cham: Springer International Publishing, 2023, pp. 333–352. doi: 10.1007/978-3-031-16203-9_20.

[6] D. Chumachenko, O. Sokolov, and S. Yakovlev, 'Fuzzy recurrent mappings in multiagent simulation of population dynamics systems', IJC, pp. 290–297, Jun. 2020, doi: 10.47839/ijc.19.2.1773.

[7] S. Dolgikh, 'Modeling of Small Data with Unsupervised Generative Ensemble Learning', CEUR-WS.org, vol. 3302, pp. 35–43, 2022.

[8] O. Mishchuk and R. Tkachenko, 'Expansion of Neural-like structures inputs using combined approximation', in Proceedings of the 3rd International Scientific Conference Computer and Information Systems and Technologies (CSITIC-2019), Kharkiv, Ukraine, Apr. 2019, pp. 29–30.

[9] W. A. Omta et al., 'Combining Supervised and Unsupervised Machine Learning Methods for Phenotypic Functional Genomics Screening', SLAS Discovery, vol. 25, no. 6, pp. 655–664, Jul. 2020, doi: 10.1177/2472555220919345.

[10] I. Tsmots, R. Tkachenko, V. Teslyuk, Y. Opotyak, and V. Rabyk, 'Hardware Components for Nonlinear Neuro-like Data Protection in Mobile Smart Systems', in 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine: IEEE, Nov. 2022, pp. 198–202. doi: 10.1109/CSIT56902.2022.10000636.

[11] H. Elghazel and A. Aussem, 'Unsupervised feature selection with ensemble learning', Mach Learn, vol. 98, no. 1–2, pp. 157–180, Jan. 2015, doi: 10.1007/s10994-013-5337-8.

[12] U. Markowska-Kaczmar and T. Switek, 'Combined Unsupervised-Supervised Classification Method', in Knowledge-Based and Intelligent Information and Engineering Systems, Springer, Berlin, Heidelberg, 2009, pp. 861–868. doi: 10.1007/978-3-642-04592-9_107.

[13] N. Shakhovska, I. Izonin, and N. Melnykova, 'The Hierarchical Classifier for COVID-19 Resistance Evaluation', Data, vol. 6, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/data6010006.

[14] Practical machine learning with python: a problem-solver's guide to building real-world intelligent systems, 1st edition. New York, NY: Springer Science+Business Media, 2017.

[15] Y. Wang, Y. Xu, and T. Gao, 'Evaluation Method of Wind Turbine Group Classification Based on Calinski Harabasz', in 2021 IEEE 5th Conference on Energy Internet and Energy System Integration (EI2), Taiyuan, China: IEEE, Oct. 2021, pp. 2630–2635. doi: 10.1109/EI252483.2021.9713300.

[16] D. S. Kim, 'Finding the Number of Clusters Using a Small Training Sequence', IEEE Access, vol. 11, pp. 25932–25940, 2023, doi: 10.1109/ACCESS.2023.3257163.

[17] 'Heart Attack Prediction'. https://www.kaggle.com/datasets/imnikhilanand/heart-attack-prediction (accessed Aug. 16, 2023).

[18] V. Kotsovsky, F. Geche, and A. Batyuk, 'Finite Generalization of the Offline Spectral Learning', in 2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP), Aug. 2018, pp. 356–360. doi: 10.1109/DSMP.2018.8478584.

Intelligent Expert Systems Application for Structuring Educational Materials to Enhance the Quality of Learning Outcomes

Alla Ivanyshyn
Department of Information Measuring Technologies
Lviv Polytechnic National University
Lviv, Ukraine
https://orcid.org/0000-0002-3302-7889

Tetiana Zahorodniuk
Preparatory Department
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
https://orcid.org/0000-0003-1060-8987

Valentyna Maliarenko
Institute of Biology and Medicine
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
https://orcid.org/0000-0003-4585-3114

Bohdan Sus
Institute of High Technologies
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
https://orcid.org/0000-0002-2566-5530

Sergiy Zagorodnyuk, Oleksandr Bauzha
Faculty of Radio Physics, Electronics and Computer Systems
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
https://orcid.org/0000-0003-3415-7746

Oksana Boyko
Department of Medical Informatics
Danylo Halytsky Lviv National Medical University
Lviv, Ukraine
https://orcid.org/0000-0002-8810-8969

***Abstract* — Application of intelligent expert systems for structuring and automation of laboratory activities in the educational process has been analysed. It is shown that combination of teaching materials with cross-references and transitions by the means of expert systems allows educators to make optimal decisions on time, set priorities and prepare more effective laboratory classes. The implementation of expert systems for classification sections of academic disciplines in both engineering and technical sciences and humanities has been discussed. An algorithm for making decisions by an expert system, where the main emphasis is given to laboratory tasks and work with laboratory equipment, is proposed. It is shown how to develop a new expert system which can help university educators prepare remote laboratory work, and also to use of virtual simulators or effective practical training. The proposed expert systems can be used for classifying thematic sections of a wide range of disciplines, including natural sciences, engineering, and humanities. They can be used for preparing new courses and training new educators in different areas. Thus expert systems are described as high-scale software units without restrictions on the depth of the question tree and the number of logical branches of the classifier.**

***Keywords* — Intelligent Expert System; Remote Laboratory Work; Virtual Laboratory Work; Virtual Simulator.**

I. Introduction

Representatives of educational institutions are forced to widely use the means of structuring and automation in the educational process [1] to ensure its competitiveness and compliance with modern standards of development [2]. Such tools include automated learning systems, electronic reference books and textbooks [3], tools for assessment and control of knowledge [4]. High-tech support of the educational process allows educators to prepare and conduct classes at a high level. In particular, students of one specialization can study related disciplines, which often contain common topics, elements, results, and conclusions. On the contrary, for students of different specialties, the same discipline may contain different material, or the same material, the topics and issues of which are presented in different sequences. As a result, it has different hierarchical nesting and structure. That formulates the need to organize methodological material in the form of a nested hierarchical structure. The article demonstrates that this technical problem can be solved using intelligent expert systems [5].

The expert system can record teaching materials prepared by one educator and, through cross-references and transitions, combine them with the teaching materials of another educator, as well as provide access to such mutually integrated materials to other fellow educators [6]. Expert systems can be applied to solve a variety of multi-purpose problems [7] in different fields of science [8], technology [9,10], business [11], medicine [1,12-15]. They allow to conveniently formalize the experience and traditions of the educational process. The training material can be saved in the form of a hierarchically structured database. Different levels of access can be arranged for professionals with different experiences. If a new educator appears among the educators of the educational institution, he or she can be assigned the mode "Read-only", which allows him or her to quickly get acquainted with the educational material. Later, such educator can become a supervisor and author of new educational material.

Thanks to the recommendations and conclusions of the expert system, educators can prepare and conduct practical, seminar, or laboratory classes more effectively.

II. Development and Configuration of the Expert System's Composition and Layout

Educators possess various approaches to generate and populate expert systems. These expert systems, once established, serve as tools for students to engage in laboratory assignments, oversee virtual simulations, and facilitate hands-on sessions in subjects like history and other humanities. Undoubtedly, preparing and configuring a virtual lab demands substantial time investment from

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

teachers. Additionally, the procedural explanation for laboratory tasks can substantially differ between scientific, technical, and IT fields. Nevertheless, during the phase of lab preparation, the expert system can empower instructors to timely and effectively make well-informed choices, establish priorities, and allocate focus.

During the phase of conceptualizing and constructing an ex-pert system, educators have the opportunity to integrate desired interface components for interlinking with diverse fields of study, replace outdated and ineffective elements, and categorize practical tasks. The majority of expert systems are accessible through web applications, compatible with contemporary operating systems such as GNU Linux, Microsoft Windows, Apple macOS, and Google Android. For instance, when working with the widely-used Exsys Corvid intelligent system, the utilization of the Oracle Java runtime environment and the Apache Tomcat web server becomes imperative. This facilitates remote connectivity and sustained utilization for other users of the expert system. Apache Tomcat and Oracle Java are cross-platform tools, thereby capable of functioning across various operating systems.

Users have the capability to engage with decentralized, multi-tiered menus that have been set up by the creator of the expert system. The expert system has an elaborate classification question hierarchy. Users are required to respond to the query at the present tier of the classification menu before progressing to the subsequent level, where the expert system poses inquiries of greater precision and specificity. Upon successfully answering all queries, the expert system generates and presents its conclusive output. At each tier of the classification, it is the responsibility of the expert system editor to either present a fresh interactive question to the user or articulate an expert decision.

The formulation and derivation of the expert system's conclusion can vary based on the subject matter. However, regardless of the field, the conclusion's content should strive to encompass a comprehensive amount of information.

III. Expert System for creating remote laboratory works

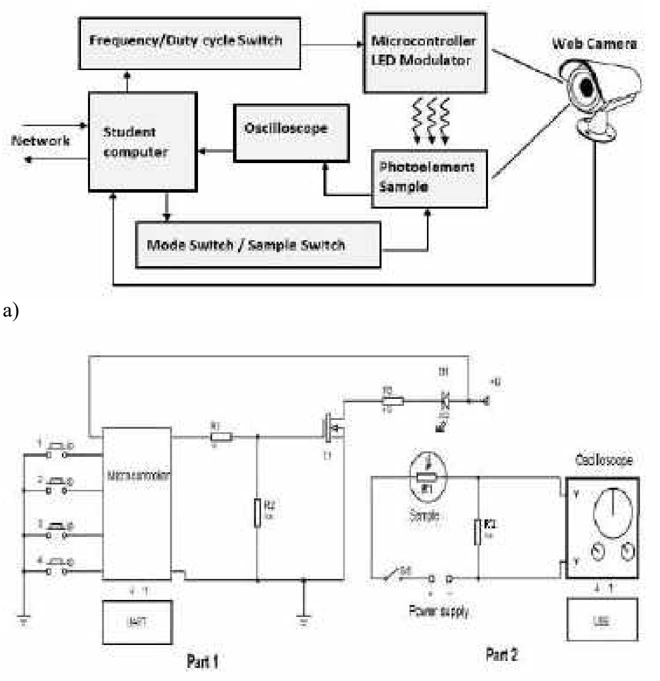
Currently, given the prolonged pandemic and war, the need to establish remote laboratories and facilitate remote laboratory works has become particularly pertinent.

Remote laboratories encompass dedicated hardware and software computing setups that afford students the chance to perform experiments being physically outside the educational institution. This involves interacting with real devices and tools, leveraging its pre-existing network infrastructure.

As an example, a teacher of the discipline "Semiconductor electronics" can choose the thematic section of the classifier of the expert system "Study of nonequilibrium charge carriers in semiconductors." In this case, the expert system can formulate a detailed description of remote laboratory work with a complete list of necessary equipment, electrical schematic diagrams, photography, network infrastructure configuration, expected results. Therefore, the conclusion of the expert system, which contains recommendations for the creation of laboratory work, can be as follows:

Remote laboratory work "Study of Mechanisms of charge carriers nonequilibrium in semiconductors".

The practical experimentation relies on a research setup devised by the authors of the article. This setup is intended for generating light pulses with a strobe-like pattern, directed towards the specimens under examination. The block diagram and electrical circuit diagram of this research apparatus are illustrated in fig. 1a and 1b.



a)

b)

Fig. 1. Block diagram (a) and Circuit diagram (b) of the remote laboratory work "Study of Mechanisms of charge carriers no equilibrium in semiconductors".

In order to assemble such an experimental setup, the following necessary hardware components must be prepared:

1. Semiconductor photocell sample.

2. N76E003AT20 microcontroller as a square-wave generator and ATMEGA328 microcontroller as RGB LED wavelength switch for sample illumination.

3. Digital software oscilloscope OWON VDS1022I or its analogue. It must be visually observed and receive saw-like pulses with the possibility of visual observation and measurement of relaxation time.

4. Webcam that allows students to monitor the progress of the measurement remotely.

An example of the entire assembled and functioning experimental setup is presented in Fig. 2.

In addition to the description of the assigned remote laboratory work, the conclusion of the expert system may also contain links to related laboratory activities developed by teachers of the same scientific, didactical or external resource with additional recommendations and alternative approaches. Different embedded systems can be used [16].

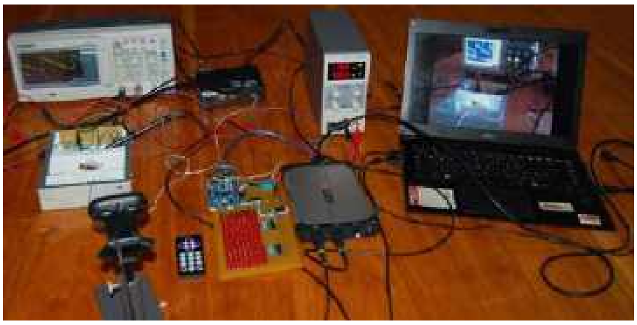


Fig. 2. Video monitoring of the laboratory setup.

IV. Expert System for preparation of virtual laboratory works

Education across numerous engineering fields necessitates the completion of various laboratory tasks [2, 3], as well as diverse projects as part of coursework. A substantial portion of these assignments and projects can be carried out through the utilization of simulators [17] or via remote network access to authentic apparatus [12]. In the former scenario, the student's workstation should be outfitted with dedicated software applications to facilitate modeling and executing essential computations. In the latter scenario, a robust Internet connection is imperative to enable remote management of the studied object. This aspect holds particular significance within the domains of natural sciences, engineering, and high-tech disciplines.

Specifically, the authors have created an array of virtual laboratory assignments centered around the "Computer Logic" field. In its entirety, the curriculum encompasses 12 laboratory works, 9 of which students complete autonomously employing the widely used circuit software, Proteus Design Suite. Each of these laboratory works entails the creation and simulation of an electronic circuit, leveraging diverse pre-existing components within the Proteus program's library. These components range from basic LEDs to intricate integrated circuits or microcontrollers. All significant elements identified by the option number can be selected by the educator during the expert system's classifier phase.

For example, the educator can choose such library elements of the Proteus software simulator as ATTiny44 microcontroller, RS-232 serial interface, 8×8 LED matrix or a set of individual LEDs, 74HC565 shift register or seven-segment indicators. Consequently, the summing-up of the expert system can begin with the following tasks:

Task 1: “Connect a COMPIM port, a DS2430 permanent storage device and eight LEDs to the ATTiny44 microcontroller. Organize the ability to execute the following commands:

a) Write one byte of data to a storage device, where the address to be written and the data byte is obtained from the serial port.

b) Reading a byte of data from the storage device and outputting it to eight LEDs. ".

Task 2: “Connect a COMPIM serial port, a DS2430 permanent storage device, and an 8 × 8 LED matrix to the ATtiny44 microcontroller. Provide the ability to execute the following commands:

a) Write an eight-byte image to a storage device, where the byte values and the address for are obtained from the serial port.

b) Read the value of the image bytes from the storage device and output the image to the 8 × 8 LED matrix. The input address is obtained from the serial port”.

After generating the task for the student, the expert system generates an example of the correct execution of the virtual simulation for this task for the educator. This example includes a Proteus circuit synthesis file, a program GUI program code that writes and receives data from the serial port, a description of the entire system, and a list of relevant and related virtual jobs. Figure 3 shows an example of correct execution of virtual laboratory work on the above task 1: synthesized circuit in Proteus, GUI-application for writing/reading data through the emulator serial interface RS-232 and output of this data to a display of LEDs.

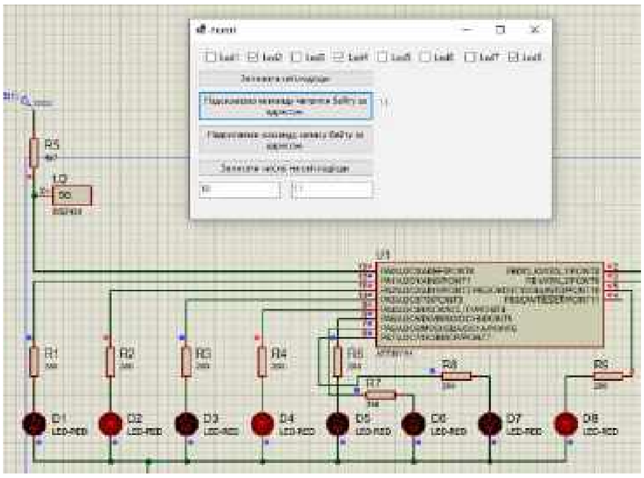


Fig. 3. Synthesized circuit in Proteus and GUI-application (Ukrainian interface) for writing/reading data through the emulator serial interface RS-232 and output of this data to a display of LEDs.

The 8×8 LED matrix contains 64 LEDs. It is quite difficult, expensive, and unreasonable to control these 64 LEDs individually using any digital circuit or microcontroller. Therefore, multiplexing is required to interact with the matrix through a minimum number of contacts. The anodes of all the LEDs in each column are connected, the cathodes of all the LEDs in each row are also connected. In this way, you can control the LED matrix through only 16 pins. In other words, it is convenient to connect the LED matrix to the microcontroller using two shift registers 74HC595. An example of the correct execution of virtual laboratory work on task 2, is presented in fig. 4.

Numerous scientific resources incorporate virtual laboratory exercises across the realms of natural and engineering subjects [18]. A widely recognized platform in this context is the "Physics Education Technology" (PhET) virtual laboratory environment [19]. The PhET program contains a collection of over 100 models, spanning a variety of disciplines including physics, mathematics, chemistry, and computer science. It offers the means to conduct visual experiments and facilitate the creation of novel virtual laboratory experiences.

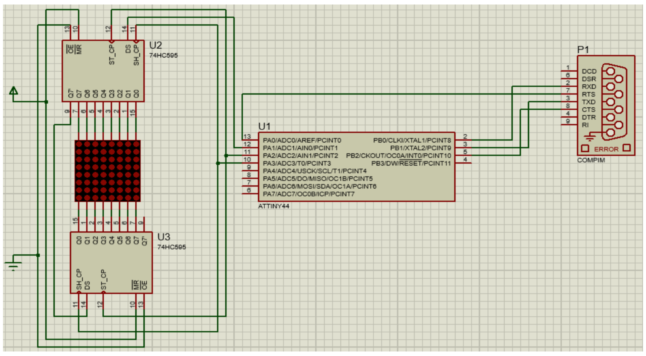


Fig. 4. Synthesized circuit in Proteus as an example of the correct execution of virtual laboratory work for task 2.

The article's authors have designed a set of virtual laboratory exercises focused on employing physical research methods to analyse the composition of chemical compounds. This assortment of laboratory tasks has been fashioned into a web-based application accessible to users globally [20].

An alternative powerful and adaptable approach to arranging virtual laboratory activities involves the utilization of containerization technologies. Notably, the widely used Docker platform serves as an example of such technologies. [21]. It permits students to virtualize not the entire operating system but exclusively a distinct service or program, enabling configuration in alignment with their specific laboratory work version. Naturally, this strategy presumes a foundational understanding of containerization on the user's part. However, initiating and engaging with Docker containers, similar to laboratory setups, is uncomplicated. Docker containers can be launched, halted, and suspended repeatedly, showcasing swift deployment. The use of Docker containers is efficient in terms of CPU and RAM resources, so they can be deployed simultaneously on cloud services or educator's computer [22].

The Crocodile ICT project has been launched to study and consolidate IT disciplines. It allows explain the construction of software algorithms and practical mechanisms for their application easily and clearly [23]. Open Educational Resources (OER): Simulations and Virtual Labs comprise a robust compilation of cross-platform Java-based virtual laboratories. These labs can be downloaded by students and operated on any contemporary operating system. [17].

V. Expert System for Lessons Preparation in the Humanities

Expert systems can also find utility in educating humanities instructors, particularly those specializing in history, philosophy, philology, and law. Interestingly, the fundamental principles of formulating and constructing expert systems remain largely consistent, even though the structuring of educational methodologies within each of these fields naturally possesses its distinctive characteristics.

In the realm of humanities education, teachers have the capacity to construct multiple concurrent classifiers, each tailored to a distinct criterion. Nonetheless, the user's navigation through these diverse classifiers, owing to their interconnectedness and cross-references, may ultimately result in the derivation of identical conclusions. For instance, the study of the history of any given nation can take various routes based on the specialization of students.

Traditionally, history is studied in chronological order. In this case, the expert system generates a series of analyses concerning the different historical periods of the nation, while refraining from delving excessively into the intricacies of the outcomes.

On the other hand, history can be studied by classifying eras and periods of history according to the causes and effects of military conflicts. It is convenient for military educational institutions.

During the study of the history of Ukraine by students in the field of law, diplomacy, statehood, the facts of the formation, dissolution or transformation of certain state institutions, unions, associations, federal and confederation ties can be emphasized.

The last level of the classifier on the history of Ukraine can look like this:

1. Name of the state and form of government of the state of Ukraine

1.1. Greek ancient polises: Chersonese, Pontic Olbia, Thira, Pantikapaion (800-200 AD)

1.2. Scythian state (700-100 AD)

1.3. Gothic state (200-400)

1.4. Establishment of Kyiv (400-500)

1.5. Kyivan Rus' Land (800-1300)

1.6. Kingdom of Galicia–Volhynia (1200-1400)

1.7. Grand Duchy of Lithuania (1400-1600)

1.8. Ukrainian lands as part of Polish–Lithuanian Commonwealth (1600-1700)

1.9. State of Bohdan Khmelnytsky "Zaporizhian Cossack Army" (1700-1750)

1.10. Ukrainian lands in the Austrian and Russian empires (1750-1917)

1.11. Ukrainian People's Republic (1917-1918)

1.12. Ukrainian State (1918)

1.13. Directory of the Ukrainian People's Republic (1918-1920)

1.14. Ukrainian Socialist Soviet Republic (1919-1937)

1.15. Ukrainian Soviet Socialist Republic (1937-1991)

1.16. Ukraine (1991-2023)

The expert system's conclusion should be crafted in a succinct manner, relying solely on essential statements and facts required to maintain the comprehensive essence of the conclusion. The teacher should focus on the pivotal circumstances of the most significant outcomes and repercussions within the historical era. This concise conclusion aids the teacher in swiftly recollecting the primary phases of the highlighted period, facilitating the identification of connections to unfamiliar details and documents.

CONCLUSIONS

1. Expert systems can organize and systematize the knowledge and experience of educators of educational institutions to ensure a quality level of assignment preparation and conduct of the educational process. Such regulation and structuring can be performed according to different criteria. Therefore the topics and sections of one discipline can be designed in the form of several parallel classifiers. On the other hand, passing through the question tree of different classifiers can lead the user to one conclusion. Expert systems are effective for classifying laboratory and practical activities. Based on the selected subject or section within the discipline, the expert system could recommend to the educator whether to organize remote laboratory exercises utilizing actual equipment or through a virtual simulation.

2. The performance of laboratory tasks within virtual simulators adheres to a specific structured procedure. The outcome of this procedure is the attainment of the learning objective. Numerous laboratories works focusing on a single topic, along with a substantial array of variations, can be amalgamated during the design phase into a unified data processing and decision-making model, delineated by a shared expert system classifier. The conclusion of the expert system should contain a sample of the correct execution of virtual laboratory work, as well as control questions with links to related virtual simulations.

3. Expert systems can be used for classifying thematic sections of a wide range of disciplines, including natural sciences, engineering, and humanities. These systems are high-scale software units without restrictions on the depth of the question tree and the number of logical branches of the classifier.

REFERENCES

[1] S. Hossain, D. Sarma, R.J. Chakma, W. Alam, M. M. Hoque, I.H. Sarker. "A Rule-Based Expert System to Assess Coronary Artery Disease Under Uncertainty" Computing Science, Communication and Security, vol.1235, pp.143–159, 2020.

[2] W. Villegas-Ch, M. Román-Cañizares, X. Palacios-Pacheco. "Improvement of an Online Education Model with the Integration of Machine Learning and Data Analysis in an LMS," Appl. Sci. vol.10, p. 5371, 2020.

[3] H. Li, W. Cui, Z. Xu, Z. Zhu, M. Feng, "Yixue Adaptive Learning System and Its Promise on Improving Student Learning," Proc. of the 10th Int. Conf. on Computer Supported Education. Setúbal Portugal, 2018, pp.45–52.

[4] W. Villegas-Ch, A. Mera-Navarrete, J. García-Ortiz, "Data Analysis Model for the Evaluation of the Factors That Influence the Teaching of University Students," Computers, 2023, vol.12, p.30. https://doi.org/10.3390/computers12020030.

[5] L Yufei, S. Saleh, H. Jiahui, S.M. Syed, "Review of the Application of Artificial Intelligence in Education," International Journal of Innovation, Creativity and Change, vol.12(8), pp.1-15, 2020. https://doi.org/ 10.53333/IJICC2013/12850.

[6] B. Sus, S. Zagorodnyuk, O. Bauzha, V. Maliarenko, T. Zahorodniuk. "Development of Practical Exercises in Educational Institutions Using Intelligent Expert Systems," Proc. of 8th International Conference on Problems of Infocommunications, Science and Technology, Kharkiv, Ukraine, 2021, pp.279–284.

[7] A.J. Paul, "Randomised fast no-loss expert system to play tic-tac-toe like a human," Cognitive Computation and Systems, vol.2(4), pp.231–241, 2020. https://doi.org/10.1049/ccs.2020.0018

[8] A. K. Akanbi, M. Masinde, "Towards the Development of a Rule-Based Drought Early Warning Expert Systems Using Indigenous Knowledge," Proc. of Int. Conf. on Advances in Big Data, Computing and Data Communication Systems, Durban, South Africa, Aug. 2018, pp.1–8, 10. https://doi.org/ 109/ICABCD.2018.8465465.

[9] I. Merino, J. Azpiazu, A. Remazeilles, B. Sierra, "2d Image Features Detector and Descriptor Selection Expert System," Proc. of 8th International Conference on Natural Language Processing, 2019, pp.51–61, https://doi.org/10.5121/csit.2019.91206.

[10] K. Crockett, A. Latham, N.Whitton, "On Predicting Learning Styles in Conversational Intelligent Tutoring Systems using Fuzzy Decision Tree," International Journal of Human-Computer Studies, vol.97, pp.98-115, 2016.

[11] M. Bohlouli, N. Mittas, G. Kakarontzas, T. Theodosiou, L. Angelis, "Competence assessment as an expert system for human resource management: A mathematical approach," Expert Systems with Applications, vol.70, pp. 83–102, 2017.

[12] I. Almarashdeh et al., "Real-Time Elderly Healthcare Monitoring Expert System Using Wireless Sensor Network," SSRN Journal, vol.13, pp. 3517-3523, 2018. https://doi.org/10.2139/ssrn.3415732

[13] H.Y. Agizew, "Adaptive Learning Expert System for Diagnosis and Management of Viral Hepatitis," IJAIA, vol.10(2), pp. 33–46, 2019.

[14] S. Mojrian et al. "Hybrid Machine Learning Model of Extreme Learning Machine Radial basis function for Breast Cancer Detection and Diagnosis; a Multilayer Fuzzy Expert System," Proc. of 2020 RIVF International Conference on Computing and Communication Technologies, Ho Chi Minh, Vietnam, 2020, pp.1–7.

[15] D. Santra, J.K. Mandal, S.K. Basu, S. Goswami, "Medical expert system for low back pain management: design issues and conflict resolution with Bayesian network," Med Biol Eng Comput, vol.58(11), pp.2737–2756, 2020. https://doi.org/10.1007/s11517-020-02222-9.

[16] H. Barylo, O Boyko, I. Helzhynskyy, R. Holyaka, "Embedded system for supply voltage converter of organic lightemitting diode with extended functionality," Przegląd Elektrotechniczny, vol. 97 (12), pp.68-72, 2021. https://doi.org/10.15199/48.2021.12.11.

[17] Open Educational Resources: Simulations And Virtual Labs. URL: https://libguides.mines.edu/oer/simulationslabs.

[18] B. Sus, N. Tmienova, I. Revenchuk, O. Bauzha, S. Stirenko, "Gamification approach to the creation of virtual laboratory works and educational courses," CEUR Workshop Proceedings, vol.2711, pp.68–78, 2020.

[19] Interactive Simulations for Science and Math: 806 million simulations delivered. https://phet.colorado.edu.

[20] A. Alam, A. Mohanty, "Design, Development, and Implementation of Software Engineering Virtual Laboratory," A Boon to Computer Science and Engineering (CSE) Education During Covid-19 Pandemic," Proc. of the 3rd International Conference on Sustainable Expert Systems, Lalitpur, vol. 587, pp. 1 - 20, 2023,

[21] A.K. Yadav, M.L. Garg, M.N. Ritika, "Docker containers versus virtual machine-based virtualization," Advances in Intelligent Systems and Computing, vol.814, pp. 141-150, 2019.

[22] S. Saito, S. Fujita, "Realtime Physics Simulation of Large Virtual Space with Docker Containers," Proc. of 22nd International Conference on Parallel and Distributed Computing, Applications and Technologies, Guangzhou, vol.13148, pp.249 – 260, 2022.

[23] R. Jamshidi, I. Milanovic, "Building Virtual Laboratory with Simulations," Computer Applications in Engineering Education, vol.30 (2), pp. 483 – 489, 2022.

Goal-Driving Control as a Base Model of the Feeling Artificial Intelligence

Anatolii Kargin
Department of Information Technology
Ukrainian State University of Railway Transport
Kharkiv, Ukraine
kargin@kart.edu.ua

Tetyana Petrenko
Department of Information Technology
Ukrainian State University of Railway Transport
Kharkiv, Ukraine
petrenko_tg@kart.edu.ua

***Abstract* — When creating unmanned systems (US), the main attention is paid to the problem of autonomy. The use of artificial intelligence (AI) is one of the ways to increase the level of US autonomy in a disordered environment. Today, a new generation of Feeling AI (FAI) is aimed to support the process of control of the implementation of action plan of Autonomous Intelligence US (AIUS). The peculiarity of the AIUS control task is that making decision in real time about current action use the state of the plan's implementation, the current situation and the ability to implement the remaining part of the action plan to achieve the goal. The article examines sequence control and rules-based control methods. The structure of a multi-layer distributed fuzzy logic system (FLS) combined with production rules system is given. A modified fuzzy inference engine which thanks to the introduction of a fuzzy certainty factor as a complex number is able to uniformly process both linguistic variables of FLS and facts from production rules system, was considered. An example of AIUS control task and computer experiments with a wheeled robot are given.**

Keywords* — *feeling artificial intelligence, autonomous intelligent unmanned system, fuzzy logic control system, production rules system, complex fuzzy certainty factor

I. Introduction

When creating new generations of unmanned cars [1], autonomous vehicles for military purposes [2] and other autonomous systems, the main attention is paid to the problem of increasing their level of autonomy [3]. Industrial US consists of automatic lines and machines with numerical control program system to which transport and robotic systems are connected [4]. In the railway industry, USs are being created on the basis of smart trains [5]. There are three classes of US depending on the level of autonomy: programmed automatic US, intelligent US and AIUS [4, 6, 7]. The first type has limitations: US can only work as pre-programmed and cannot adapt to any changes in the environment. The second type has some perception, decision-making and control capabilities, and can adjust itself according to changes in the environment. AIUS has a high level of autonomy, can autonomous decision-making in wide range of uncertainty [8]. AI is an important component that ensures this level of autonomy.

US hardware resources set the framework for autonomy, and the intelligence provided by AI technologies determines the level of decision-making autonomy in this frame. When developing smart things, there is not focused on. However, to create US, three types of artificial intelligence models that support the autonomy of US are discussed - mechanical, thinking, and feeling AI [7, 9]. Now the FAI is relevant for AIUS and various aspects from social consequences to architectural projects are discussed [8, 10].

FAI, as a model of a new generation AI, is designed to support the implementation of the US mission in conditions of uncertainty. This purpose of AI imposes certain requirements on its model. First, the FAI must function in the on-line mode: receiving data from sensors and decisions-making in real time based on the stream of heterogeneous multimodal data from sensors and implementing control decisions through the US actuators. Secondly, FAI resources are involved for decisions-making which support autonomous mode of US operation. Thirdly, FAI decision which was making in autonomous mode, should not harm the US environment, especially people. Last two features are conflicted therefore FAI should have extra facilities to find a compromise for the safe resolution of this contradiction. The above listed significantly distinguish the FAI model from another direction of AI development, namely artificial general intelligence [11, 12]. FAI architecture blueprint base on knowledge granularity concept is proposed [13, 14]. FAI carries out processing of the stream of data from sensors. From this point of view, it is presented by three layers: perception, decision-making and control, and action implementation. The main component of perception layer is Cognitive Perception System (CPS) which organizes as multi-level structure represented knowledge “What Is This” type. Results of processing are sense of spatial-temporal segment of data from sensors obtained by abstraction in the space of granules of data from sensors and convolution of data stream in time [15, 16]. Decision-making and control layer represents knowledge “How Do It” type. Result of processing on this layer is degree of relevance of stage of action plan to current segment of data from sensors. Action implementation layer organizes as multi-level structure, too. Knowledge granule of *i*th level represents sequences of actions of according level of generalization. At this layer, results of processing are states of AIUS actuators [13, 14].

In this paper the model of decision-making layer is discussed. At first, based on analysis of problem of control in AIUS there are discussed restrictions of "pure" sequence control and rules-based control models. Then we introduce modified fuzzy control model which rely on specificity of plan implementation control. And finally, we describe prototype of decision-making and control layer system of FAI.

II. Problem of Plan Implementation Control

The first step when creating US is to present its mission as an action plan leading to a global goal. The FAI is intended to solve the problem of implementing the mission of the AIUS in the face of various obstacles. To be capable to do this, FAI must have needs to overcome these obstacles. At any stages of implementation of mission plan, the obstacles can arise. FAI should localize this situation, activate the appropriate need which intended to overcome these obstacles, and switch control from mission's plan to need's plan. After elimination

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

of obstacles, the FAI should turn back control to the interrupted stage of mission plan for continues its implementation. Thus, FAI carries out management of plans and separate stages of actual plan, and controls an action for achievement the local goal of actual stage. FAI decision making is a multi-stage process in real time. At each stage of decision-making, the current state of plan implementation, the current situation, and the possibility of completing the remaining part of the action plan to achieve the goal are taken into account.

For the clarity this problem we will discuss on the following simple example of US mission. Let the mission of a wheeled robot, as a type of AIUS, consist in the continuous movement of cargo from position A to position B (Fig. 1). The mission implementation plan is represented by the sequence of states of the "Environment-AIUS" system, which the AIUS, thanks to its actions of actuators, must pass through in sequence. This plan can be represented as a sequence of states (1).

$$(B^{Load},\ 4^{Mov},\ 1^{Mov},\ A^{Mov},A^{UnLd},\ \ldots,3^{Mov},\ 5^{Mov},\ B^{Mov}) \quad (1)$$

In (1), individual stages of the plan are represented by local goals, for example, A^{Mov} is a state "AIUS has arrived at position A". Three needs for overcoming three types of obstacles are considered. Energy replenishment is a first need when low battery charge requires a visit to a charging station. Self-preservation is the second need when an object-obstacle is encountered in the path of the AIUS movement or destruction of the environment (the marks on the floor are worn or there is no side fence). Environment scrutiny is the third need when data for decision making or knowledge to identify situation are absent. Each of these needs may be presented by actions plan as a sequence of local goals (1) leading to the elimination of obstacles.

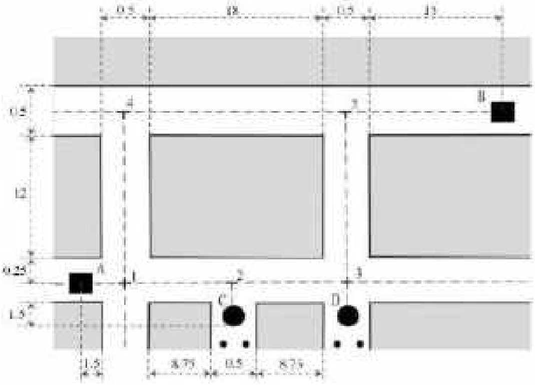


Fig. 1. AIUS environment model.

III. Sequence Program Control or Rule-Based Control: What is Better for Goal-Driving Control

A. Sequence Program Control Methods

Two types of sequence control are used: direct and feedback. Sequence control methods are widely used to control the robots and another industry US of first class mentioned in introduction [17-19]. In addition, the majority of intelligent control algorithms developed for second class of US belong to this type, too [20, 21].

To use the sequence control methods in decision of discussed problem, the actions plan (1) is divided into separate stages (frames), the sequence of which represents the control program of AIUS. The engine manages the control program frame by frame: it checks the completion of the stage and makes a decision on the proceeding to the next stage. In the direct sequence control method, the condition of the stage completion is time (the time interval is specified in special command "*delay x*" in the frame). For plan (1) of Fig. 1 environment, the required time for moving robot with speed *v=3* m/sec between precisely specified points, for example, $S_{B\to 4}$ = 33.75m is calculated. It is $t_{B\to 4}$ = 10.25 sec. This value is specified in the command "Delay" in Table I. Except command "Delay" control program in Table I used else three commands "Go_ahead", "Left" and "Right" with "on" or "off" parameters. In the feedback control method, data from sensors, and not time intervals, are used to identify the conditions of completion of the plan stage.

TABLE I. Examples of Control Program of AIUS

Sequence Control Methods		
Direct control	***Feedback control***	***Intelligent control***
Go_ahead on Delay $t_{B\to 4}$=10.25 Go_ahead _off Left on Delay $t_{90°}$ = 0.5 Left off Go_ahead on Delay $t_{4\to 1}$=3.75 Go_ahead _off Right Delay $t_{90°}$ = 0.5 Right_off Go_ahead on Delay $t_{1\to A}$ = 0.45 Go_ahead _off	Go_ahead 33.75 Left 90 Go_ahead 12.5 Right 90 Go_ahead 1.5	{1st frame: (MovF 0.5), (MoveV 12.0), (Go_ahead 33.75)} {2nd frame: (Left 90)} {3rd frame: (MovL 0.25), (MoveV 6.0), (MovM M$^+$), (Go_ahead 12.5)} {4th frame: (Right 90)} {5th frame: (MoveV 4.0), (MovM M+), (MovL 0.25), (Go_ahead 1.5)}

In second column of Table 1, shows the control program in which the commands use numerical parameters. The condition for completion of the "Go_ahead" command is set by a value of the traveled distance, for example, $S_{B\to 4} = 33.75$ in Table I. Similarly, the completion of the turn commands ("Left", "Right") is set by a value of the rotation angle, for example, *γ=90°*. On the basis of feedback control method, various options of implementation plan control systems are created, for example, intelligent control, where it is possible to set the stage completion condition of any complexity. The control program there may include stages with different control methods, that is, a stage of direct control may be followed by a stage that implements feedback control, when the condition of stage completion is the distance traveled, or the distance to an obstacle, or the presence of a certain marker identified by a vision camera. In third column of Table I, shows the intelligent control program in which module names with their parameters are used: "MovL" is module that controls the motion along the marking lane on the floor, which corrects the deviation from the marker of line; "MovF" is module that controls movement along the artificial fence at a given distance from it; "MoveV" is module that controls the motion velocity; "MoveM" is module that localize neither the the robot is at given marker, for example, intersection sign (M$^+$), or not.

The direct control method is demanding for the prior arrangement of the environment: mandatory requirements to the constant speed of movement and instant acceleration/deceleration of the robot. It is almost never possible to fulfill these requirements for wheeled robots.

Feedback control methods have advantages over direct control methods, due to the use of sensor data [17]. The possibilities of autonomous performing the plan are expanding due to use the actual data, and not on expected preliminary calculations. Despite the indicated advantages, feedback control methods, including intelligent control, have disadvantages that limit their ability to support AIUS autonomy. Namely, disturbances lead to situations (state of the environment) when the implementation of the action according to the stage becomes impossible: there is action, but there is no expected result. For example, in order to achieve the goal, it is necessary for the AIUS to move from the position where it is to position 5. In the frame of control program, the goal is set as a condition for the completion of the action, and not the specified conditions that are required for the activation of the command to lead to the successful implementation of the move. This contradicts the principles of autonomy and is a limitation for the "pure" use of all the sequence control methods in AIUS discussed above.

B. Rule-Based Control Methods

In robotics, the Internet of Things, and smart machines, the following rule-based artificial intelligence models are used to make control decisions in real time based on various data, including from sensors [22]:

- Rule-based systems with symbolic representation of knowledge and symbolic inference engine (Rule-based Production System, RbPrS).
- Rule-based systems with symbolic representation of knowledge and probabilistic inference engine (Bayesian models).
- Rule-based systems with symbolic representation of knowledge and certainty factor inference engine.
- Rule-based systems with linguistic variables knowledge representation and fuzzy inference engine (Fuzzy Logic System, FLS).

All four rule-based models listed above satisfy the requirements for the decision-making, taking into account certain conditions specified in the rules [23]. The first type in form as classical production rules system has limitation due to requirements of knowledge completeness and absence of any uncertainty [23, 24]. The next two models based on the probabilistic and certainty factor inference engines take into account the incompleteness of knowledge related to cause-and-effect relationships, and the decision is made only in conditions of such uncertainty. They do not take into account another kind of uncertainty, namely the vagueness or fuzziness of the objects themselves, which are mentioned above in cause-and-effect relationships. At the same time, FLS processing both types of uncertainties, consequently from this point of view has advantages. However, the engines of first three model can work the sequences of plan stages, since they do multi-step sequential derivation of intermedial local goals before getting a global goal. Thus, if combine two FLS and RbPrS models, in advance overcome limitation each of them, then such integrated AI model will satisfy requirements of plan implementation control system.

The above-mentioned limitations are following. A fuzzy system is problematic, at first, to set up if number of FLS inputs exceed 5-7, and, at second, to tune it when adding new input numerical variables or changing the terms of linguistic variables [25]. There is FLS problem known as dimension's problem. Forward chaining and backward chaining inference model of RbPrS can't works with knowledge incompleteness and uncertainty [22]. There is RbPrS uncertainty's problem. If above two problems to solve, then integrated fuzzy rules-based system FLS&RbPrS can be used as goal-driving control model, which is keeping the advantages both the FLS based on linguistic variables and forward and backward chaining inference engine of RbPrS [22].

IV. Goal-Driving Control Model

A. Goal-Driving Fuzzy Control System Structure

The FLS large dimension problem can be overcome thanks to the peculiarity of the controlling process of the plan implementation. The peculiarity is that not all rules or groups of rules simultaneously take part in determining the control decision, but only those that are related to the implementation of the actual stage of the plan. The large dimension problem is overcome according the principle "Divide and conquer". Problem is divided into separate independent subtasks of controlling the individual stages of the plan. For this, the Knowledge Base (KB) is structured. The set of rules is divided into separate Local KBs (LKBs) according to the stages of the plan. Similarly, such LKBs are created for the stages of needs plans, which are responsible for controlling the implementation of these plans in case of various disturbances. These Local FLSs (LFLSs) are independent of each other.

The RbPrS uncertainty's problem can be overcome thanks to introduced universal model of fuzzy certainty factor on base of which the Fuzzy RbPrS (FRbPrS) is created, namely fuzzy representation of facts and rules and fuzzy engine model of FRbPrS. On the base of FRbPrS, the continues planning engine [20, 24] which tracking stages of plan is possible to realize.

The structure of FAI decision-making layer on the base of FLS&RbPrS consists of the components of both types of systems: the FLS and FRbPrS (Fig. 2). FLS is represented by such traditional components as FLS Engine and FLS KB. The latter is formed by local knowledge bases LKB_1, LKB_2,…, LKB_n.

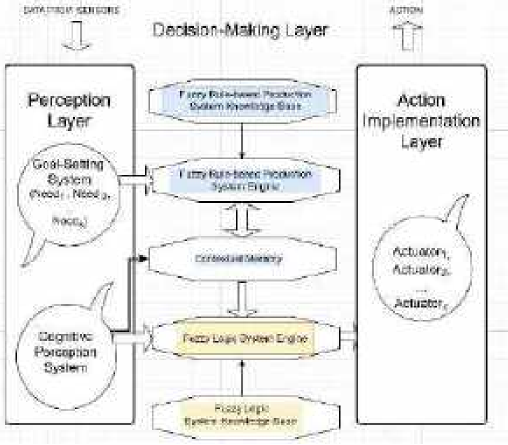

DATA FROM SENSORS
Decision-Making Layer
ACTION
Perception Layer
Goal-Setting System (Need1, Need2, … NeedN)
Cognitive Perception System
Fuzzy Rule-based Production System Knowledge Base
Fuzzy Rule-based Production System Engine
Contextual Memory
Fuzzy Logic System Engine
Fuzzy Logic System Knowledge Base
Action Implementation Layer
Actuator1, Actuator2, … Actuatork

Fig. 2. Structure of FAI decision-making layer.

The FRbPrS is also represented by components traditional for production systems: FRbPrS KB, FRbPrS Engine, and the fact base, the role of which is played by Context Memory (CntxMem). The latter contains a set of facts used to represent the state of implementation of the stages of possible action plans. The numerical input variables of the FLS are the

Presumed Certainties (PCs) of the data from the CPS and facts from the CntxMem. The output numerical variables of FLS are the PCs of the AIUS control actions. The FLS Engine uses fuzzy rules with linguistic variables from FLS KB when deriving control actions. The input numerical variables of FRbPrS are PCs of facts from CntxMem and Goal-Setting System. The latter contains the facts the PCs of which characterizes the levels of relevance of different needs of AIUS. The output numerical variables of FRbPrS are the PCs of facts in CntxMem. The FRbPrS Engine changes the PCs values of the facts in CntxMem based on the knowledge from FRbPrS KB about the sequence of plan stages. The latter consist of k local KBs about action plans for AIUS needs, including the mission. As can be seen from Fig. 2, facts from CntxMem are used by both fuzzy systems FLS and FRbPrS, and only FRbPrS can change the values of the PCs of facts in CntxMem. In Fig. 2, double arrows denote the flow of facts PCs, single arrows denote the flow of knowledge from knowledge bases, and the dotted arrow denotes the relationship of the PCs values of paired facts in CPS and CntxMem.

B. Presumed Certainty are Quantitative Assessment of Senses of the Fact and Linguistic Variable

The PC is defined based on fuzzy certainty factor which is a fuzzy LR number **X** with a Gaussian L-R membership function [14]

$$\mathbf{X}:\{x \mid m_{\mathbf{X}}(x), \forall x \in [-q,+q],\ q \geq +1\} \tag{2}$$

where

$$m_{\mathbf{X}}^{L}(x)=\exp(-(x-\alpha)^2/2\cdot(v_L\cdot t_L)^2),\ \forall x\in[-1,\alpha]$$
$$m_{\mathbf{X}}^{R}(x)=\exp(-(x-\alpha)^2/2\cdot(v_R\cdot t_R)^2),\ \forall x\in(\alpha,+1];$$

$-1.0 \leq \alpha \leq +1.0$ is certainty; t_L is the time interval that has passed since the moment of receiving the data; t_R is the time interval that has passed since the data change; v_L, v_R are data aging rate coefficients.

The PC is a crisp number which corrects the certainty α according to the aging of data. Defuzzification model of fuzzy LR number (2) is below.

$$cf=\alpha\cdot k_t \tag{3}$$

where

$$k_t=1-\frac{\sum\limits_{\forall x\in[-1,\alpha)} m_{\mathbf{X}}^{L}(x)+\sum\limits_{\forall x\in(\alpha,+1]} m_{\mathbf{X}}^{R}(x)}{Card([-1,+1])-1}.$$

The aging of data over time leads that the confidence estimates $cf \approx +1$ or $cf \approx -1$ tends to zero $cf \approx 0$, which characterizes uncertainty (complete lack of confidence). For cases when time intervals are small (or from the moment of receiving data, or, in special cases, from the moment of data change), the confidence does not change much compared to α, that is, $cf \approx \alpha$.

Based on the expert's knowledge, the CPS distills the sense of data from sensors and presents it by a set of facts, namely by the fuzzy certainty factors and PCs of facts. The FLS Engine uses only PCs as numerical FLS inputs. In addition to these, FLS Engine uses PCs of facts from context memory as an input's numerical variables, too (Fig. 2). The numerical outputs of FLS Engine which transfer to the actuators are PCs, too. All linguistic variables of FLS are homogeneous determined on the universe of PC *(–1.0 ≤ cf ≤ +1.0)* by three terms *low*, *un* and *high* certainties with trapezoidal membership functions *m = (a, b, c, d)* as show below: m_{low}= *(−1.0, −1.0, −0.75, −0.25)*; m_{un}=*(−0.75, −0.25, 0.25, 0.75)*; m_{high} = *(0.0, +0.4, +1.0, +1.0)*.

C. Features of FLS

The processing of distributed FLS KBs requires modification of the FLS engine. First, at each moment of time, the engine must "know" which of the LFLS is relevant now. The second feature is related to the implementation of an action. Action is "triggered" by an activated rule when a certain event occurs [26]. The activated state of rule lasts for a certain time until a local goal is achieved, and during this time interval, the rule state changing is inadmissible. To take this feature into account when processing rules, it is necessary to be able to temporarily "hide" certain LKB rules, and the engine should be able to "see" them only when certain events specified in the rule occur. The third feature that helps tracking the sequence of stages is the ability to management the context during the implementation of the plan. For this, it is necessary to distinguish the contextual facts from CntxMem.

Below are examples of rules in which shown all the features listed above.

R$_0$: ***if*** ***Event***(cf_$^*1^{Mov}$ is high) ***and*** cf_4^{Mov} is high ***and***
cf_M$^-$ is high ***and*** cf_P is high
then cf_MovL is high, V is high
R$_1$: ***if*** cf_$^*1^{Mov}$ is high ***and Event***(cf_M^0 is high) ***and***
cf_D^R&Close is high ***and*** cf_M$^-$ is low
then cf_MovL is low, cf_MovF is high, V is middle
R$_2$: ***if*** cf_$^*1^{Mov}$ is high ***and Event***(cf_M$^-$ is high)
then cf_MovL is high, cf_MovF is low, V is high
R$_3$: ***if*** cf_$^*1^{Mov}$ is high ***and Event***(cf_M$^+$ is high) ***and***
cf_1^{Move} is high
then cf_Stop is high, V is zero (4)

Rules (4) belong to the LFLS that controls the implementation of the 1^{Mov} stage. An asterisk in the name of linguistic variable, for example, cf_$^*1^{Mov}$ indicates that this is a contextual fact. Rule, for example R$_0$, will be triggered (activated) only once when the local goal of the actual stage of plan appears ***Event***(cf_$^*1^{Mov}$ *is high*). It is happened, when the execution of the previous stage has completed, which is indicated by the fact cf_4^{Mov} *is high*. For all other times, this rule is hidden from the FLS engine and is not processed. If the AIUS CPS has localized markings on the floor *(*cf_M$^-$ *is high)*, then the moving along lane (cf_M$^-$ *is high*) with high speed (*V is high*) is activated (cf_MovL *is high*). Rule R$_1$, will be activated once, too when the event "violated marking on the floor" occurs, along which the robot moves to position *1*. This is indicated by the ***Event***(cf_M^0 *is high*). The presence of an artificial side fence (cf_D^R&Close *is high*) is needed to deactivates the module of control motion along the lane and activates the module of control of motion along the railing. Rule R$_2$ returns to the module of control of moving along the lane when it reappears. When local goal of current stage of plan is reached (the CPS locates the position number marker (cf_1^{Mov} *is high*) and the AIUS is above the floor intersection marker ***Event***(cf_M$^+$ *is high)*), the R$_3$ rule is activated, which deactivates the previous control cf_Stop *is high.* As can be seen in (4), the rules from the LKB are activated only if the local goal of the stage is activated (cf_$^*1^{Move}$ *is high*).

Thus, in order for the Mamdani FLS engine can to implement fuzzy inference according to the rules (4), it must be modified as follows. The relevance of the event specified in the rule by the ***Event***() function is checked first. An event is considered relevant when t_R = 0, where t_R is the time parameter of the fuzzy certainty factor in (2). When inferring, the FLS engine uses only rules with actual events and rules that do not contain references to events.

D. Features of FRbPrS

As mentioned above, at each stage decision is making only when remained part of plan will be implemented successfully. Control with continuously replanning today is focus in US [20, 26-28]. Such approach is done in FLS&RbPrS: FLS supports control function and FRbPrS carry out replanning. The FRbPrS consists of three subsystems (Fig.2). The role of fact base of traditional production model performs CntxMem. As already mentioned above, the state of the AIUS environment in CPS is represented by a set of sensor's facts, the certainty factor of which is determined by current values of sensors. Let CPS contain a set of facts Ω^{CPS}, which is divided into two subsets Ω^{plan} and Ω^{sit}. The first contains facts used to define the plan (1), and the second contains facts describing the situation surrounding AIUS.

$$\boldsymbol{\Omega}^{CPS} = \{\boldsymbol{\Omega}^{plan} = \{f_j, j=1,2,\dots,n\} = \{A^{UnLd}, B^{Load}, 1^{Mov}, 2^{Mov}, 3^{Mov}, 4^{Mov}, 5^{Mov}, A^{Mov}, B^{Mov}\}, \boldsymbol{\Omega}^{sit} = \{g_j, j=1,2,\dots,m\}\}. \quad (5)$$

For each fact f_j from Ω^{plan}, the CntxMem contains a context fact $^*f_j$ to track the plan's execution stage and map it to the current context. We denote the contextual facts $^*f_j$ as well as sensory f_j in (5) only with the superscript star. Facts f_i and $^*f_j$ are paired: changing of PC value of f_i causes a change in the state of paired fact $^*f_j$ in CntxMem. This relation is reflected by a single arrow in the Fig. 2. Thus, CntxMem contains set of paired facts.

$$\boldsymbol{\Omega}^{CntxMem} = \{^*f_j, j=1,2,\dots,n\} = \{^*A^{UnLd}, {}^*B^{Load}, {}^*1^{Mov}, {}^*2^{Mov}, {}^*3^{Mov}, {}^*4^{Mov}, {}^*5^{Mov}, {}^*A^{Mov}, {}^*B^{Mov}\}. \quad (6)$$

The state of the paired contextual facts $^*f_j$ is determined by Complex PC (CPC). The CPC is complex fuzzy number $z_j = a_j + b_j i$, where a_j and b_j are the real numbers of PC (3) $-1.0 \le a_j, b_j \le +1.0$. The real part of CPC $Re(z_j) = a_j = cf^{Re}$ is certainty factor of paired fact from CPS. Its value is changed by CntxMem engine when value of paired fact in CPS has changed. The imaginary part of CPC $Im(z_j) = b_j = cf^{Im}$ is changed by FRbPrS engine when carry out the planning. At any time, the state of plan implementation is presented by the state of the CntxMem.

$$\textbf{State} = \{z_j = cf^{Re} + cf^{Im}i, j=1,2,\dots,n\}. \quad (7)$$

The following states of context fact are possible. The state $z_j, = -1.0 + 1.0i$ describes that fact $^*f_j$ is current local goal, state $z_j, = +1.0 + 1.0i$ describes that fact $^*f_j$ is previously achieved local goal, and $z_j, = cf^{Re} + 0.0i$, $|cf^{Re}|<1.0$ reflect that fact $^*f_j$ was the goal which was achieved some times ago. For example, at current time AIUS has arrived at position *5* from position *B* where it has been loaded and according to plan (1) next local goal is position *4* (Fig. 1). The following state of CntxMem describes this situation.

$$\textbf{State}=\{z_4 = -1.0+1.0i,\ z_5 = +1.0-1.0i,\ z_{*BLoad} = +0.75+0.0i,\ \{z_j = 0.0 + 0.0i,\quad j={}^*A^{UnLd}, {}^*1^{Mov}, {}^*2^{Mov}, {}^*3^{Mov}, {}^*A^{Mov}, {}^*B^{Mov}\}\}. \quad (8)$$

The CntxMem engine, when activated, carry out simple operation. First, since the current goal has been achieved, engine changes the state of corresponding fact $^*f_j$ by multiplication of $z=cf^{Re}+cf^{Im}i$ by $-1.0i$.

$$z_j = (-1.0 + 1.0i)\cdot(-1.0i) = (1.0 + 1.0i) \quad (9)$$

Second, there are changing the state of fact $^*f_j$ that describes the previously achieved local goal by operation of subtract.

$$z_j = (1.0 + 1.0i) - (0.0 + 1.0i) = (1.0 + 0.0i). \quad (10)$$

Third, aging data of all another contextual fact by multiplication of $cf^{Re}+cf^{Im}i$ by coefficient of "forgetting" $exp(-v)$, where $0 \le v \le 1$ is the data aging rate [16].

$$z_j = (cf^{Re} + 0.0i)\cdot exp(-v) = exp(-v)\ cf^{Re} + 0.0i \quad (11)$$

The FRbPrS KB contains the experience of AIUS in the form of fragments of the trajectory of behavior that led to the achievement of the goal in the implementation of the mission or the elimination of obstacles. These fragments represent, among other things, various options for the implementation of action plans. Each fragment represents the structure of knowledge in the form of Fig. 3.

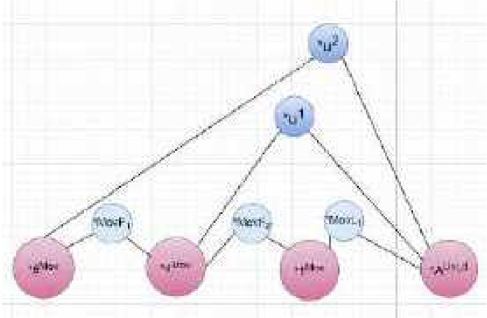


Fig. 3. Structure of knowledge portion of FRbPrS KB.

At the bottom level of the fragment structure, elementary knowledge is given in the form of "$fact_1$-act-$fact_2$", which reflect that in the presence of $fact_1$, the action *act* leads to the appearance of $fact_2$. Below in (12) are a set of rules that represent a fragment of the structure of Fig. 3. The structure reflects not only elementary causal relationships, but also stable more longer relationships chains. For example, goal $^*A^{UnLd}$ is achieved (the fact $^*A^{UnLd}$ can be obtained) if, in the presence of fact $^*5^{Mov}$, action $^*u^2{}_1$ is implemented, which is a generalization of the sequence of actions $^*MovL_1$ and $^*u^1{}_1$. Similarly, goal $^*A^{UnLd}$ is achieved if, in the presence of fact $^*4^{Mov}$, action $u^1{}_1$ is implemented which is a generalization of the sequence of actions $^*MovL_2$ and $^*MovL_3$. This knowledge is represented in the knowledge base by R₃ and R₄ rules.

R₀: ***if*** $^*1^{Mov}$ ***and*** $^*MovL_1$ ***then*** $^*A^{UnLd}$ *(cf=0.9)*
R₁: ***if*** $^*4^{Mov}$ ***and*** $^*MovF_1$ ***then*** $^*1^{Mov}$ *(cf=0.9)*
R₂: ***if*** $^*5^{Mov}$ ***and*** $^*MovF_2$ ***then*** $^*4^{Mov}$ *(cf=0.9)*
R₃: ***if*** $^*5^{Mov}$ ***and*** $^*MovF_3$ ***then*** $^*3^{Mov}$ *(cf=0.9)*
….
R₄: ***if*** $^*5^{Mov}$ ***and*** $^*u^2{}_1$ ***then*** $^*A^{UnLd}$ *(cf=0.6)*
R₅: ***if*** $^*4^{Mov}$ ***and*** $^*u^1{}_1$ ***then*** $^*A^{UnLd}$ *(cf=0.75)* (12)

The FRbPrS engine implements inference mechanism with combination of forward and backward chaining inference techniques. The forward chaining inference uses the real parts of CPC of facts from CntxMem. Before the activation of the FRbPrS engine, the real parts of the facts CPCs in the CntxMem describe the real state of the environment and the stages of the plan, since the CntxMem engine maps the state of the facts from CPS into the cf^{Re} values of the CntxMem facts. The CntxMem engine maps the states of not only the facts describing the stage of the plan implementation, but also the CPS facts characterizing the possibility of implementing actions leading to the achievement of local goals. For example, $cf^{Re} = 0.9$ of the fact $^{*}MovL_1$ reflects the possibility of realizing a move from $^{*}1^{Mov}$ to $^{*}A^{UnLd}$. Another value of $cf^{Re} = -0.9$ indicates the impossibility of achieving a local goal $^{*}A^{UnLd}$ by this way (rule R_0 in (12)). These values change according to changes in the state of the environment in real time. In the case of backward chaining inference techniques (from goal fact to the fact that reflects the real state of the implementation of the plan, i.e. the completed stage of the plan), the imaginary parts of CPCs cf^{Im} are used. The initial state of CntxMem before the start of FRbPrS engine is prepared by the Need engine (Fig. 2) by $cf^{Im} = 0.9$ for facts describing the global goal, for example, $^{*}A^{UnLd}$. After the completion of the FRbPrS engine operation, the fact that has the maximum value of the CPC module $|cf^{Re}+cf^{Im}i|$ and argument $0 \le \arg(z) \le 90°$ is found. This fact represents current local goal used by the FLS when processing the rules (4).

Conclusion

A prototype of system FLS&RbPrS has been created. The structure of Fig. 2 has multileveled organization in which the traditional feedback controllers are located on the first level, the sequence control modules are located on the second level and the goal-driving fuzzy control model are implemented on third level. The computer experiments were conducted on the example of the task considered in this article. FLS LKBs and FRbPrS KB were created for three types of hazards, namely damage of marking, low battery, and obstacles in the robot's path, and also for mission of AIUS and the need of scrutiny of the environment. Designing the traditional FLS aimed to control the implementation of an action plan for autonomous cargo robot taking into account above conditions, characterized by more than 40 input numerical variables from sensors, is an insoluble task. Proposed model FLS&RbPrS which consists of the components of both types of systems FLS and FRbPrS overcome all problems and saves advantages of FLS in handling uncertainty.

In the future, it is planned to develop a prototype of imbedded system with a multilayer structure based on microcontrollers with a set of sensors.

References

[1] L. Joseph, A.K. Mondal, Eds, Autonomous Driving and Advanced Driver-Assistance Systems (ADAS). Applications, Development, Legal Issues, and Testing. 1st edn. CRC Press, Boca Raton, 2021, doi:10.1201/9781003048381.

[2] “Sikorsky and DARPA's Autonomous Black Hawk Flies Logistics And Rescue Missions Without Pilots On Board,” Lockheed Martin Corporation, 2022, USA, Accessed: Aug. 10, 2023. [Online]. Available: https://news.lockheedmartin.com/2022-11-02-Sikorsky-and-DARPAs-Autonomous-Black-Hawk-R-Flies-Logistics-and-Rescue-Missions-Without-Pilots-on-Board

[3] J. Deichmann et al., “Autonomous driving’s future: Convenient and connected,” McKinsey Center for Future Mobility. Report, Jan. 2023, Accessed: August 15, 2023. [Online]. Available: https://www.mckinsey.com/

[4] H. Chen et al., “From Automation System to Autonomous System: An Architecture Perspective,” J. of Marine Sci. and Eng., vol. 9, no. 6, Jun. 2021, doi: 10.3390/jmse9060645.

[5] Rail Technical Strategy. Innovating across Britain’s railway. Oct. 2022. Accessed: Aug. 10, 2023. [Online]. Available: https://railtechnicalstrategy.co.uk/wp-content/uploads/2022/10/The-Rail-Technical-Strategy.pdf

[6] T. Zhang et al., "Current trends in the development of intelligent unmanned autonomous systems," Frontiers Inf. Technol. Electron. Eng., vol. 18, Feb. 2017, pp. 68–85, doi: 10.1631/FITEE.1601650.

[7] J. Reis, Y. Cohen, N. Melao, J. Costa, and D. Jorge, "High-Tech Defense Industries: Developing Autonomous Intelligent Systems," Appl. Sci. , vol. 11, 4920, 2021, doi: 10.3390/app11114920.

[8] J. Chena, J. Sun, and G.Wang, "From Unmanned Systems to Autonomous Intelligent Systems," Engineering, vol.12, May 2022, pp. 16-19, doi: 10.1016/j.eng.2021.10.007.

[9] M. Czerwinski, J. Hernandez, D. Mcduff, "Building an AI that feels," Appl. Sci., vol.11, 4920, Apr. 2021, doi: 10.3390/app11114920.

[10] M. Huang and R. Rust, "Artificial Intelligence in Service," J. of Service Res., vol. 21(2), Feb. 2018, pp.155-172, doi: 10.1177/1094670517752459.

[11] G. Singer, "Thrill-K: A Blueprint for The Next Generation of Machine Intelligence," Towardsdatascience.com. Accessed: Aug. 15, 2023. [Online]. Available: https://towardsdatascience.com/thrill-k-a-blueprint-for-the-next-generation-of-machine-intelligence-7ddacddfa0fe

[12] J. Bach, Principles of Synthetic Intelligence: Psi: An Architecture of Motivated Cognition, 1st ed. Oxford University Press, 2009.

[13] A. Kargin, T. Petrenko “Feeling Artificial Intelligence for AI-Enabled Autonomous Systems,” in Conf. Proc. 2022 IEEE Global Conf. Artif. Intell. and Internet of Things (GCAIoT), Alamein New City, Egypt, Dec. 2022, pp. 88-93.

[14] A. Kargin, T. Petrenko “Knowledge Distillation for Autonomous Intelligent Unmanned System,” in W. Pedrycz, S-M. Chen, Eds., Advancements in Knowledge Distillation: Towards New Horizons of Intelligent Systems. Studies in Computational Intelligence, vol. 1100. Springer International Publishing, 2023, pp. 193-231.

[15] A. Kargin and T. Petrenko, “Multi-level Computing With Words Model to Autonomous Systems Control,” in Proc. 9th Int. Conf. Inf. Control Sys.&Tech (ICST-2020), A. Pakstas, T. Hovorushchenko, V. Vychuzhanin, H. Yin, N. Rudnichenko, Eds, Odessa, Ukraine, Sep. 2020, CEUR Workshop Proceedings, vol. 2711, pp. 16-30.

[16] A. Kargin and T. Petrenko, “Method of Using Data from Intelligent Machine Short-Term Memory in Fuzzy Logic System,” in Conf. Proc. of 2021 IEEE 7th World Forum on Internet of Things (WF-IoT), New Orleans, Louisiana, USA, Jun. 2021, pp. 842-847.

[17] K. Michels, F. Klawonn, R Kruse, A. Nurnberger, “Fundamentals of Control Theory,” in Fuzzy Control. Fundamentals, Stability and Design of Fuzzy Controllers, 2006, Springer, Heidelberg, pp. 57-234.

[18] R. Babuska, E.A. Mamdani, Fuzzy control. Scholarpedia 3(2): 2103, 2008, doi:10.4249/scholarpedia.2103.

[19] J. Yu, “Research on mobile robot path planning and tracking control,” Inter. J. of Computational Sci. and Eng., vol. 26, no. 4, Jul. 2023, pp. 349-360, doi: 10.1504/IJCSE.2023.132164.

[20] Y. Hu et al., “Planning-oriented Autonomous Driving,” Mar. 2023, doi:10.48550/arXiv.2212.10156.

[21] M. Leonetti, L. Iocchi, P. Stone, “A synthesis of automated planning and reinforcement learning for efficient, robust decision-making,” Artificial Intelligence, vol. 241, Dec. 2016, pp. 103-130, doi:10.1016/j.artint.2016.07.004.

[22] J. C. Andersen, O. Ravn, N. A. Andersen, “Autonomous Rule-Based Robot Navigation in Orchards,” IFAC Proceedings Volumes, vol. 43, no. 16, 2010, pp. 43-48, doi: 10.3182/20100906-3-IT-2019.00010.

[23] M. Negnevitsky, Artificial Intelligence: A Guide to Intelligent Systems. 2nd ed. Addison-Wesley, 2005.

[24] S.J. Russell and P. Norvig, Artificial Intelligence. A Modern Approach, 3rd ed. Upper Saddle River, NJ, USA: Pearson Education, 2010.

[25] A. Piegat, Fuzzy modelling and control. Heidelberg: Physica-Verlag Heidelberg New York, 2001.

[26] J.R. Sanchez-Ibanez, C.J. Perez-del-Pulgar, A. Garcia-Cerezo, “Path Planning for Autonomous Mobile Robots: A Review,” Sensors 2021, 21, 7898, doi:10.3390/s21237898.

[27] S. Trimpe and J. Buchli, “Event-based estimation and control for remote robot operation with reduced communication,” 2015 IEEE Int. Conf. Robot. & Automat. (ICRA), doi: 10.1109/ICRA.2015.7139897.

[28] I. Gokasar et al., “Metaverse integration alternatives of connected autonomous vehicles with self-powered sensors using fuzzy decision making model,” Information Sciences, vol. 642, Sep. 2023, doi: 10.1016/j.ins.2023.119192.

Development of a Ray Tracing Framework for Simulating Acoustic Waves Propagation Enhanced by Neural Networks

Oleksandr Terletskyi
Applied Mathematics and Informatics
Ivan Franko National University of Lviv
Lviv, Ukraine
oleksandr.terletskyi@lnu.edu.ua

Petro Venherskyi
Applied Mathematics and Informatics
Ivan Franko National University of Lviv
Lviv, Ukraine
petro.venherskyy@lnu.edu.ua

Valeriy Trushevskyi
Applied Mathematics and Informatics
Ivan Franko National University of Lviv
Lviv, Ukraine
valeriy.trushevsky@lnu.edu.ua

Ostap Hrytsyshyn
Applied Mathematics and Informatics
Ivan Franko National University of Lviv
Lviv, Ukraine
ostap.hrytsyshyn@lnu.edu.ua

***Abstract* — Accurate simulation of sound propagation is pivotal across a spectrum of disciplines ranging from virtual reality to architectural acoustics. This study delves into the innovative convergence of ray tracing techniques and neural networks to improve the existing implementations of sound propagation simulation. Traditional methods, though valuable, are hindered by computational demands and inherent approximations. In response, this future research is going to introduce a novel approach that harnesses the precision of ray tracing and the pattern recognition prowess of neural networks.**

***Keywords* — Framework, Machine Learning (ML), Sound Propagation Simulation, Ray Tracing, Acoustic Modelling**

I. Introduction

The problem we are addressing in this research is the limited accuracy and efficiency of traditional methods for simulating sound propagation in complex environments. Traditional approaches, such as wave-based and ray-based methods, struggle to capture intricate interactions, diffraction effects, and complex sound paths [1]. This limitation becomes particularly pronounced in applications like virtual reality, architectural acoustics, and entertainment, where realistic sound propagation is crucial for creating immersive and engaging experiences [3].

Consequently, there is a growing need for a simulation approach that can accurately model sound interactions, including reflections, diffractions, and occlusions, while also maintaining computational efficiency. This problem is further exacerbated by the increasing demand for real-time and interactive simulations in fields like virtual reality, where both realism and responsiveness are paramount.

To address this problem, our research explores the integration of ray tracing techniques with neural networks. By leveraging the precision of ray tracing and the pattern recognition capabilities of neural networks, we aim to develop a novel approach that overcomes the limitations of traditional methods. This approach has the potential to enhance the accuracy of sound propagation simulations, enable the accurate modelling of diffraction and complex interactions, and significantly improve computational efficiency.

Accurate sound propagation simulation holds significant importance in various fields due to its potential to enhance user experiences, improve design processes, and enable realistic virtual environments [2]. Here are some key areas where accurate sound propagation simulation is crucial:

A. Virtual Reality (VR) and Augmented Reality (AR)

In immersive virtual environments, realistic audio plays a vital role in creating a convincing sense of presence. Accurate sound propagation simulation helps in delivering spatially precise and coherent audio cues, enhancing the immersion and overall quality of VR and AR experiences.

B. Architectural Acoustics

Architects and designers rely on accurate sound propagation simulation to predict how sound will behave in a given space. This is crucial for designing auditoriums, concert halls, classrooms, and other architectural spaces where optimal acoustics are desired for speech clarity, music performance, and overall comfort.

C. Entertainment Industry

In film, gaming, and multimedia production, accurate sound propagation contributes to a more immersive and engaging experience for audiences. Realistic audio enhances storytelling, adds depth to scenes, and enhances emotional engagement.

D. Industrial Design

Industries that involve noise-sensitive processes or equipment benefit from accurate sound simulation to design quieter environments, thereby improving worker safety and comfort.

E. Environmental Noise Assessment

Urban planners and regulators use sound propagation simulation to assess the impact of proposed developments on the surrounding noise environment. This helps in designing noise mitigation strategies and adhering to noise regulations.

F. Communication Systems

Accurate sound propagation simulation is essential in designing and optimizing communication systems like public address systems, emergency announcements, and

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

teleconferencing setups to ensure clear and intelligible speech.

II. Ray Tracing in Acoustics

A. Principle of Ray Tracing

The principle of ray tracing is a fundamental concept in computer graphics and simulation that involves tracing the path of rays as they interact with objects in a scene. Originally developed for rendering realistic images, ray tracing has found applications in various fields, including acoustics simulation [5]. Here's a concise explanation of the principle of ray tracing:

In ray tracing, the process begins with the emission of rays from a virtual camera or an audio source. These rays represent paths that light or sound energy might take as it travels through a scene. As rays propagate, they interact with objects in the environment, such as surfaces and materials. The interactions can include reflection, refraction, diffraction, absorption, and transmission.

Key steps in the ray tracing process:

- Ray Generation: Rays are cast from a virtual camera or sound source into the scene. For sound propagation, these rays simulate the paths that sound waves would take from a source to a receiver.

- Intersection Testing: Each ray's path is traced through the scene to determine if it intersects with any objects, such as walls, floors, or obstacles. This is typically done using geometric calculations.

- Surface Interaction: When a ray intersects an object's surface, various interactions can occur. In graphics, this includes calculating lighting and shading for visualisation. In acoustics, this involves understanding how sound waves interact with surfaces, including reflections, diffraction, and absorption.

- Secondary Rays: Additional rays can be generated as a result of interactions. For example, a ray that hits a reflective surface may generate a new ray representing the reflected path.

- Recursive Tracing: To capture complex interactions like reflections and transparency, ray tracing often involves recursion. That is, a traced ray can generate new rays that continue the simulation of energy propagation.

- Gathering and Calculation: At the endpoints of rays, information is collected, and calculations are performed. In graphics, this could involve determining the color and brightness of a pixel. In acoustics, this could involve calculating the intensity and phase of sound waves.

The principle of ray tracing provides a powerful framework for simulating how light or sound interacts with the environment, enabling the creation of realistic images or accurate sound propagation simulations. This approach offers a high level of detail and accuracy but can be computationally intensive due to the need to trace multiple rays and simulate their interactions. proceedings, and not as an independent document. Please do not revise any of the current designations.

B. Overview of Potential Challenges

Ray tracing techniques, originally developed for computer graphics, offer valuable insights into sound propagation in acoustics. However, there are certain challenges and considerations that arise when applying ray tracing to simulate sound in complex environments. Here are some of the notable challenges:

- Computational Complexity: Ray tracing involves tracing a significant number of rays through the environment and calculating their interactions with surfaces. In acoustics, this can be computationally intensive, especially for scenes with numerous reflective surfaces and complex geometries. The computational demands can limit real-time applications or require significant computational resources.

- Diffraction Modeling: While ray tracing excels at modeling reflections and specular interactions, accurately simulating sound diffraction around obstacles is more challenging. Diffraction involves the bending of sound waves around corners or obstacles, and capturing these effects accurately requires specialized algorithms that can handle wavefront interactions.

- Frequency Dependence: The accuracy of ray tracing in acoustics can be influenced by the frequency of the sound waves. Higher-frequency waves tend to be better captured by ray tracing, while lower-frequency waves may exhibit diffraction and other behaviors that are harder to simulate using rays.

- Sound Source Modeling: The accurate representation of sound sources is essential for realistic simulations. In ray tracing, representing complex sound sources with different directivity patterns or multiple simultaneous sources can be complex and may require advanced techniques.

- Materials and Absorption: Sound absorption and material properties are important factors in accurate sound propagation simulation. Incorporating the absorption and reflection characteristics of various materials into ray tracing calculations can be challenging and requires detailed material models.

- Multiple Reflections and Paths: In environments with multiple reflective surfaces, sound waves can follow complex paths involving multiple reflections. Capturing all possible reflection paths accurately with ray tracing requires tracing a significant number of rays, which can contribute to computational overhead.

- Spatial Sampling: For accurate results, ray tracing simulations often require dense spatial sampling, meaning a high density of rays is needed to accurately capture the intricacies of sound propagation. This can further increase computational demands.

- Validation and Verification: Validating ray tracing simulations for acoustics can be challenging due to the complexity of sound interactions in real-world scenarios. Ensuring that simulated results match physical measurements or other validated simulation methods is crucial.

Despite these challenges, researchers continue to advance the application of ray tracing to acoustics, developing techniques that address diffraction, optimisation, and integration with other methods. The combination of ray tracing with neural networks, as proposed in this study, is one such innovative approach that seeks to overcome some of these challenges and enhance the accuracy and efficiency of sound propagation simulations.

III. Neural Networks Implications In Audio Ray Tracing

Neural networks have shown promise in enhancing ray tracing techniques for various applications, including graphics [6] and potentially acoustics. Integrating neural networks with ray tracing can address certain challenges and improve the efficiency and realism of simulations. Here's an overview of how neural networks can be applied to ray tracing:

A. Acceleration of Ray Tracing

Neural networks can be trained to predict the outcomes of certain ray tracing calculations. For instance, they can predict which rays are likely to hit surfaces, which rays will be reflected, or the colors/intensity of pixels in graphics rendering. This prediction can help filter rays that are less likely to contribute to the final result, accelerating the simulation process.

B. Denoising

Ray tracing simulations can produce noisy images or results due to the stochastic nature of sampling rays. Neural networks can be trained to denoise these results, enhancing the final output without significantly increasing the computational cost.

C. Implicit Representations

Neural networks can learn to represent complex implicit functions, which can be useful for modeling intricate surfaces, lighting effects, or acoustic interactions. This can lead to more efficient rendering or simulation processes.

D. Upscaling and Super-Resolution:

Neural networks can be used to enhance the resolution of ray-traced images or simulations. This is particularly useful in scenarios where higher resolution is desired without significantly increasing the computational load.

E. Predictive Sampling

Neural networks can guide the selection of rays to trace, focusing on areas where significant interactions are likely to occur. This helps allocate computational resources more effectively [4].

F. Adaptive Techniques

Neural networks can be employed to dynamically adjust simulation parameters based on the scene or the behavior of the rays, leading to more efficient and accurate simulations.

G. Complex Material Models

Neural networks can learn to approximate complex material models, allowing for more accurate representation of light or sound interactions with surfaces.

H. Hybrid Methods

Neural networks can be integrated with traditional algorithms, combining their strengths. For instance, combining ray tracing with machine learning-based approaches can improve the accuracy of sound propagation simulations by accounting for complex interactions and diffraction.

I. Transfer Learning

Neural networks trained for one scenario can be fine-tuned for a related scenario, potentially reducing the amount of data needed for training.

It's worth noting that while neural networks offer many advantages, they also come with challenges like data requirements, generalization to different scenarios, and potential overfitting [6]. Research in this area is ongoing to optimize the design and application of neural networks for ray tracing in both graphics and acoustics. This research on simulating sound propagation using ray tracing with neural networks could contribute to this evolving field and provide valuable insights into the benefits and challenges of this approach.

Conclusion

The convergence of ray tracing techniques and neural networks in simulating sound propagation marks a significant advancement in the field of acoustics and its various applications. This study shows the potential of this novel approach to address the limitations of traditional sound propagation simulation methods. Through a comprehensive investigation of existing research [6, 5], the integration of ray tracing with neural networks has been revealed as a promising avenue to enhance both the accuracy and efficiency of sound propagation simulations.

The importance of accurate sound propagation simulation across domains such as virtual reality, architectural acoustics, and entertainment has been underscored. Traditional methods [1, 2], while valuable, exhibit limitations in handling complex interactions, diffraction effects, and computational demands. These challenges have paved the way for the introduction of an innovative approach that leverages ray tracing's precision with neural networks' pattern recognition capabilities.

The methodology introduced in this study harnesses the principles of ray tracing to accurately simulate sound wave interactions with surfaces. By incorporating neural networks, intricate relationships and patterns within sound propagation have been captured, leading to more realistic simulations. Through making research in this field of study, the superior accuracy and performance of this approach in light transport is showing potential to outperform conventional methods in capturing complex sound propagation phenomena.

References

[1] A. Chandak, L. Antani, M. Taylor, D. Manocha “Fast and Accurate Geometric Sound Propagation Using Visibility Computations”, 2011

[2] J. Park, (2019). Sound Propagation and Reconstruction Algorithm Based on Geometry. International Journal of Online and Biomedical Engineering (iJOE), 15(13), pp. 86–94. https://doi.org/10.3991/ijoe.v15i13.112

[3] N. Raghuvanshi, R. Narain and M. C. Lin, "Efficient and Accurate Sound Propagation Using Adaptive Rectangular Decomposition," in IEEE Transactions on Visualization and Computer Graphics, vol. 15, no. 5, pp. 789-801, Sept.-Oct. 2009, https://doi.org/10.1109/TVCG.2009.28

[4] J. F. Talbot, D. Cline, P. Egbert, “Importance Resampling for Global Illumination” in Eurographics Symposium on Rendering, 2005

[5] A. Krokstad, S. Strøm, S. Sørsdal, “Calculating The Acoustical Room Response by the Use of a Ray Tracing Technique” in Acoustical Laboratory, The Technical University of Norway, Trondheim, Norway, 1967

[6] J. Knodt, J. Bartusek, S.-H. Baek, F. Heide, “Neural Ray-Tracing: Learning Surfaces and Reflectance for Relighting and View Synthesis”, 2021. https://doi.org/10.48550/arXiv.2104.13562

Smoothed Particle Hydrodynamics Implementation Using Compute Shaders

Ostap Hrytsyshyn
Applied Mathematics and Informatics
Ivan Franko National University of Lviv
Lviv, Ukraine
ostap.hrytsyshyn@lnu.edu.ua

Petro Venherskyi
Applied Mathematics and Informatics
Ivan Franko National University of Lviv
Lviv, Ukraine
petro.venherskyy@lnu.edu.ua

Valeriy Trushevskyi
Applied Mathematics and Informatics
Ivan Franko National University of Lviv
Lviv, Ukraine
valeriy.trushevsky@lnu.edu.ua

Oleksandr Terletskyi
Applied Mathematics and Informatics
Ivan Franko National University of Lviv
Lviv, Ukraine
oleksandr.terletskyi@lnu.edu.ua

***Abstract* — Smoothed Particle Hydrodynamics (SPH) has emerged as a versatile method for simulating fluid dynamics and various natural phenomena [6]. In this paper, we present an approach to SPH simulation using compute shaders, leveraging their parallel processing capabilities to enhance performance and accuracy. Our implementation focuses on efficient data organization, neighbor search, force computation, and particle integration within the compute shader framework. We discuss the intricacies of translating the SPH algorithm into the shader domain and optimizing memory access patterns to achieve high computational throughput. Through comprehensive experimentation, we demonstrate the effectiveness of our compute shader-based SPH implementation, showcasing improved simulation fidelity and reduced computation time when compared to traditional CPU-based approaches. We also discuss the implications of our findings for real-time fluid simulations and other computational domains. Our work contributes to the ongoing exploration of compute shaders in graphics and simulation, offering insights into the potential of parallel computing for advancing SPH techniques.**

Keywords* — *Smoothed Particle Hydrodynamics (SPH), Graphics Processing Unit (GPU), Central Processing Unit (CPU), Shader Storage Buffer Objects (SSBOs), Uniform Buffer Objects (UBOs).

I. Introduction

Smoothed Particle Hydrodynamics (SPH) is a prominent numerical technique for simulating complex fluid behaviors. However, its computational demands can be a limiting factor, particularly for large-scale scenarios. Recent advances in Graphics Processing Units (GPUs) and their programming models, such as compute shaders, offer a way to overcome these limitations by leveraging parallel processing capabilities [4].

This paper aims to explore how SPH can be efficiently implemented using compute shaders on GPUs. We examine the adaptation of the SPH algorithm to this new paradigm and compare the performance and accuracy of our approach to traditional CPU-based methods. The goal is to understand how to best utilize GPUs for real-time or near-real-time fluid simulations. Through this work, we aim to contribute to the growing field of GPU-based simulations and advance the understanding of SPH's compatibility with modern parallel computing techniques.

II. Toward Next-Generation Fluid Simulations

The focus of this article culminates in the development of a high-performance fluid dynamics simulation using compute shaders, particularly using the Smoothed Particle Hydrodynamics (SPH) model. However, it is worth acknowledging that this work represents a significant stepping stone in a broader, more ambitious research agenda.

While the current project aims to optimize fluid dynamics simulations to the fullest extent possible with existing computational approaches, the long-term vision extends far beyond. The highly efficient compute shader-based SPH implementation serves a dual purpose: it acts not only as a high-performance simulation model but also as a means for generating a rich and reliable dataset for future research.

The ultimate goal is to employ these data to train neural networks that can handle fluid dynamics simulations more efficiently than even the most optimized traditional computational models, such as the compute shader-based SPH model discussed herein. While neural network-based approaches for fluid dynamics are beyond the scope of this current work, the efficient simulation model developed here does have broader implications. Not only does it provide a more streamlined method for conducting fluid dynamics simulations as they are traditionally understood, but it also offers a reliable foundation for any future research that aims to further explore and possibly enhance simulation techniques.

By elevating the computational efficiency and simulation accuracy of current models, this research paves the way for future work that can further disrupt and transform the fluid simulation landscape.

III. SPH Basics

The subsequent section will delve into the derivation of the SPH framework from its foundational principles. This discussion will elucidate the process by which a continuous field can be translated onto a discrete particle representation, thereby facilitating an approximation of the field's behavior. The section will also explore the inherent inaccuracies introduced during this approximation. Furthermore, the approach for calculating derivatives within this context will be demonstrated. Additionally, the section will explore techniques for effectively smoothing the particle arrangement to faithfully represent the underlying field.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

A. Discretized Representations of Continuous Fields

The challenge of simulating fluid dynamics computationally hinges on an essential paradigm shift: transitioning from the theoretical continuum of fluid properties to their tangible discrete counterparts. This approach allows for computational tractability and more manageable numerical operations. The foundation of this transition often rests on mathematical tools that bridge the continuous and the discrete worlds.

A primary mathematical tool employed in this context is the Dirac delta function, denoted as δ. Conceptually, the Dirac delta function can be envisioned as an infinitely sharp spike located at the origin with an integral (or area under the curve) of one. In mathematical terms, for any function $f(\mathbf{r})$, its representation using the Dirac delta function is:

$$f(\mathbf{r}) = \int f(\mathbf{r}')\delta(\mathbf{r}-\mathbf{r}')d\mathbf{r}' \quad (1)$$

Breaking this down, $f(\mathbf{r}')$ represents the function's value at some neighboring position $\mathbf{r}'$. The role of the Dirac delta function in this integral is pivotal: it ensures that the output $f(\mathbf{r})$ is influenced only by values at the precise location $\mathbf{r}$.

This is, in essence, a representation of continuity in the field. However, the Dirac delta function, while instrumental in theory, poses challenges in computational applications due to its non-compact and infinitely sharp nature. To overcome these limitations, computational practitioners often replace the Dirac delta function with something more amenable: a compact smoothing kernel, denoted as W. This kernel has a finite extent or "support" and offers a smoothed approximation of the Dirac delta function. The modified representation becomes:

$$f(\mathbf{r}) = \int f(\mathbf{r}')W(\mathbf{r}-\mathbf{r}',h)d\mathbf{r}' \quad (2)$$

In this context, $W(\mathbf{r}-\mathbf{r}',h)$ symbolizes the smoothing kernel. The parameter h, often referred to as the "smoothing length", determines the width or extent of the kernel's influence. As h approaches zero, the kernel narrows and approximates the behavior of the Dirac delta function:

$$\lim_{h\to 0} W(\mathbf{r}-\mathbf{r}',h) = \delta(\mathbf{r}-\mathbf{r}') \quad (2.1)$$

Yet, to achieve computational feasibility, especially in simulations with thousands to millions of particles or fluid elements, we must discretize this continuous representation. Thus, our continuous integral transforms into a summation over discrete elements or particles. This discretized form is expressed as:

$$\mathrm{f}(\mathbf{r_a}) = \sum_b m_b \frac{f(\mathbf{r_b})}{\rho_b}\mathrm{W}(\mathbf{r_a}-\mathbf{r_b},h) \quad (3)$$

To unpack this: $f(\mathbf{r_a})$ signifies the field value at a discrete particle or position $\mathbf{r_a}$. Each particle in the simulation, indexed by b, carries with it certain properties, such as mass m_b and density ρ_b. The summation implies that the field value at any given particle a is influenced by neighboring particles within the extent of the smoothing kernel. The term $W(\mathbf{r_a}-\mathbf{r_b},h)$ serves as a weight, modulating the contribution from neighboring particle b based on its distance to particle a and the smoothing length h.

B. Gradients Approximation

In SPH, gradients play a pivotal role in characterizing and evolving the field quantities. The formulation's accuracy and stability are primarily influenced by the accuracy of these gradient approximations.

Given our foundational representation of a field f in SPH from equation (3), we can proceed to determine its gradient at a point $\mathbf{r_a}$:

$$\nabla f(\mathbf{r_a}) = \sum_b m_b \frac{f(\mathbf{r_b})}{\rho_b}\nabla W(\mathbf{r_a}-\mathbf{r_b},h) \quad (4)$$

In this equation:

- $\nabla f(\mathbf{r_a})$ denotes the gradient of the field f at the particle position $\mathbf{r_a}$.
- $\nabla W(\mathbf{r_a}-\mathbf{r_b},h)$ is the gradient of the smoothing kernel function, which depends on the relative positions of the particles and the smoothing length h.
- Non-symmetry: The gradient formula presented here is in a non-symmetric form. While it might not have the conservation properties of its symmetric counterpart, this form can be computationally efficient in certain scenarios.
- Kernel Choice: The choice of the smoothing kernel, evident in the term ∇W, profoundly affects the precision of the gradient approximations. Different kernels possess distinct characteristics which might be more fitting for specific applications or domains.

C. Divergence Approximation

Divergence in the context of SPH quantifies the rate at which field quantities disperse from a point. It is particularly important for understanding and modeling compressible flows and other phenomena where the local density variation of particles is significant.

Using the SPH formalism and building upon our previous discussions, we can formulate the divergence of a vector field f at particle a as:

$$\nabla\cdot\mathrm{f}(\mathbf{r_a}) = \sum_b m_b \frac{\mathrm{f}(\mathbf{r_b})}{\rho_b}\cdot\nabla W(\mathbf{r_a}-\mathbf{r_b},h) \quad (5)$$

Here:

- $\nabla\cdot f(\mathbf{r_a})$ is the divergence of the vector field f at the position $\mathbf{r_a}$.
- $f(\mathbf{r_b})$ represents the vector field value at the neighboring particle b.
- The term $\nabla W(\mathbf{r_a}-\mathbf{r_b},h)$ is the gradient of the smoothing kernel function, which, as before, is determined by the relative positions of the particles and the smoothing length h.

Key observations:

- Interpretation: The divergence approximation effectively captures the net rate of outflow of the vector field f at the point $\mathbf{r_a}$. A positive divergence indicates a net outflow, while a negative divergence indicates a net inflow.

- Kernel Sensitivity: Much like the gradient approximation, the accuracy of the divergence approximation is strongly influenced by the choice of the smoothing kernel. Kernels with better spatial resolution and desirable properties can lead to more accurate divergence calculations.

D. Numerical Accuracy

Numerical accuracy remains paramount when implementing Smoothed Particle Hydrodynamics (SPH) algorithms, especially in contexts where subtle nuances can lead to significant deviations in outcomes. In SPH, one key quantity that is commonly used to assess the fidelity of the simulation is the color field value. This field, when evaluated correctly, should produce a constant value, typically 1, across the computational domain. Its consistency serves as a litmus test for the precision of the numerical approximations involved.

$$C(\mathbf{r_a}) = \sum_b \frac{m_b}{\rho_b} W(\mathbf{r_a} - \mathbf{r_b}, h) \tag{6}$$

Ideally, for any particle a the color field value $C(\mathbf{r_a})$ should evaluate to 1 everywhere. If there are deviations from this expected value, it signals potential inaccuracies in particle representation, the weighting kernel, or other aspects of the SPH method.

Further, the gradient of the color field, which gives insights into the rate of change or variance of the field across space, should be zero throughout:

$$\nabla C(\mathbf{r_a}) = \sum_b \frac{m_b}{\rho_b} \nabla W(\mathbf{r_a} - \mathbf{r_b}, h) \tag{7}$$

For a well-behaved SPH simulation, the gradient $\nabla C(\mathbf{r_a})$ should ideally evaluate to zero. Any non-zero gradients can indicate errors, inconsistencies, or boundary-related issues in the simulation.

E. Better Approximation for Gradients

Smoothed Particle Hydrodynamics (SPH) relies heavily on the accurate approximation of gradients to simulate various physical phenomena. In this section, we'll delve into a generalized formula for gradient approximations that offer flexibility and are applicable to different scenarios. This generalized formula can be particularly advantageous in simulations where we need to account for different properties like pressure, density, and other scalar fields.

Initially, let's consider a function f in combination with density ρ, where ρ is raised to some arbitrary power n:

$$\nabla(f\rho^n) = nf\rho^{n-1}\nabla\rho + \rho^n\nabla f \tag{8}$$

The generality of this formulation allows us to look at different special cases by choosing appropriate values for n. To focus solely on the gradient of the function f, we rearrange the general equation:

$$\nabla f = \frac{1}{\rho^n}(\nabla(f\rho^n) - nf\rho^{n-1}\nabla\rho) \tag{9}$$

In the SPH framework, we deal with discretized quantities. The gradient of f in this discretized world is:

$$\nabla f(r_a) = \frac{1}{\rho(r_a)^n} \sum_b m_b (f(r_b)\rho(r_b)^{n-1} - nf(r_a)\rho(r_a)^{n-1}) \nabla W(r_a - r_b, h) \tag{10}$$

For $n = 1$, the formula is beneficial for approximating gradients where the function f is largely constant:

$$\nabla f(r_a) = \frac{1}{\rho(r_a)} \sum_b m_b \big(f(r_a) - f(r_b)\big) \nabla W(r_a - r_b, h) \tag{11}$$

In cases where $n = -1$, the approximation becomes symmetric, which has significant implications in simulations involving pressure and the conservation of momentum:

$$\nabla f(r_a) = \rho(r_a) \sum_b m_b \left(\frac{f(r_a)}{\rho(r_a)^2} + \frac{f(r_b)}{\rho(r_b)^2}\right) \nabla W(r_a - r_b, h) \tag{12}$$

The symmetric nature of this formula ensures that the pressure forces between particles are equal and opposite, which is critical for the conservation of momentum in the system. This is particularly advantageous in simulations involving compressible fluids and fast-moving particles where conservation laws are pivotal.

F. Smoothing Kernels

Smoothed Particle Hydrodynamics (SPH) inherently relies on the concept of smoothing over a discrete set of particles to represent continuous fields. Smoothing kernels play a pivotal role in this approximation. They provide a weighted averaging scheme that translates discrete particle properties into a continuous distribution. This section focuses on the importance, selection criteria, and various types of smoothing kernels used in SPH simulations.

The smoothing kernel $\mathrm{W(r_a - r_b, h)}$ is essentially a weighting function. It determines how much influence a neighboring particle b has on the field value at another particle a. Here, h is the smoothing length that controls the range of influence. The fundamental properties of smoothing kernels ensure that they:

- Are normalized: $\int \mathrm{W(r - r', h)\ dr'} = 1$
- Tend to a Dirac delta function as $\mathrm{h} \to 0$
- Are symmetric: $\mathrm{W(r - r', h) = W(r' - r, h)}$

Selection of an appropriate smoothing kernel can influence the accuracy, stability, and computational efficiency of an SPH simulation. Here are some factors to consider:

- Support Size: A kernel with a small support size will be computationally less expensive but may compromise on accuracy.
- Smoothness: More derivatives of the function should exist and be continuous for better accuracy.
- Computational Cost: Some kernels are computationally more demanding due to the complexity of their functional forms.

Popular Choices of Smoothing Kernels:

- Gaussian Kernel. The Gaussian kernel is smooth and has infinite support. However, it's computationally expensive.

$$W_G(\mathbf{r},h)=\frac{1}{\pi^{3/2}h^3}e^{-\frac{\mathbf{r}^2}{h^2}}$$

- Cubic Spline Kernel. A commonly used kernel in SPH simulations, it has a compact support and is computationally efficient.

$$W_{CS}(\mathbf{r},h)=\frac{\alpha}{h^3}\begin{cases}1-\frac{3}{2}q^2+\frac{3}{4}q^3 & 0\le q<1\\ \frac{1}{4}(2-q)^3 & 1\le q<2\\ 0 & 2\le q\end{cases}$$

- Wendland Kernel. This kernel is useful for its computational efficiency and smoothness properties.

$$\mathrm{W_W(r,h)}=\frac{\alpha}{\mathrm{h}^3}\left(1-\frac{|\mathrm{r}|}{\mathrm{h}}\right)^4\left(4\frac{|\mathrm{r}|}{\mathrm{h}}+1\right)$$

Smoothing kernels serve as the cornerstone for field approximations in SPH. Their selection depends on the trade-offs between accuracy, smoothness, and computational cost. Understanding the properties and characteristics of various kernels can lead to more robust and accurate SPH simulations.

G. Fluid Equations in SPH

The fundamental equations governing fluid flow – continuity, momentum, and energy equations – are crucial for accurately modeling fluid dynamics in Smoothed Particle Hydrodynamics (SPH). This section aims to provide an understanding of how these equations are formulated and approximated within the SPH framework.

The continuity equation describes the conservation of mass and, in a differential form, can be expressed as:

$$\frac{d\rho}{dt}+\nabla\cdot(\rho\mathbf{v})=0 \tag{13}$$

In SPH, this equation can be discretized as:

$$\frac{d\rho_a}{dt}=-\rho_a\sum_b m_b\mathbf{v_{ab}}\cdot\nabla_a W_{ab}(h) \tag{14}$$

Here, $\mathrm{v_{ab}}=\mathrm{v_a}-\mathrm{v_b}$ and $\mathrm{W_{ab}(h)}$ is the smoothing kernel.

The momentum equation, also known as the Navier-Stokes equation, is given by:

$$\rho\left(\frac{d\mathbf{v}}{dt}\right)=-\nabla P+\mu\nabla^2\mathbf{v}+\rho\mathbf{g} \tag{15}$$

The SPH form can be expressed as:

$$\frac{d\mathbf{v_a}}{dt}=-\sum_b m_b\left(\frac{P_a}{\rho_a^2}+\frac{P_b}{\rho_b^2}+\Pi_{ab}\right)\nabla_a W_{ab}(h) \tag{16}$$

Here, Π_{ab} is the artificial viscosity term.

For the momentum equation, the symmetric form ensures the conservation of momentum, making it more suitable for problems where these conservation laws are essential, like in simulations involving high-pressure gradients or shocks.

H. SPH Algorithm Pseudocode

Initialization Steps:

1. Initialize all simulation parameters including the time step size, the smoothing length, particle mass, and other physical constants.

2. Assign initial positions, velocities, and densities to all the particles.

3. Define the smoothing kernel and its gradient based on the physical properties.

Main Simulation Loop

1. For every particle, calculate its neighboring particles. Neighbors are those particles that lie within a sphere of radius equal to the smoothing length around the particle.

2. Calculate the density of each particle by summing up the contributions from all its neighboring particles, weighted by the smoothing kernel.

3. Compute the intermediate velocity for each particle. This is based on its current velocity, the forces acting on it, and the time step size.

4. Solve for the pressure gradient for each particle in such a way that the rate of change of density over time is zero. This ensures that mass is conserved.

5. Update the velocity and position of each particle. The new velocity is calculated from the intermediate velocity and the pressure gradient. The new position is updated based on this new velocity and the time step.

6. Increment the simulation time by the time step.

IV. Compute Shaders for SPH Implementation

In the domain of Smoothed Particle Hydrodynamics (SPH), compute shaders offer a powerful medium for parallel computation. Unlike the typical vertex or fragment shaders, which are tied to specific stages of the graphics pipeline, compute shaders are general-purpose. This enables them to handle arbitrary computations, making them particularly useful for computationally-intensive algorithms like SPH.

The primary advantage of using compute shaders in SPH lies in their ability to perform computations in parallel. SPH is inherently a parallel algorithm; each particle's new state can be computed independently of the others, but relies on the information from its neighboring particles. This makes it an excellent fit for the parallel architecture provided by compute shaders.

Moreover, compute shaders bring an added level of flexibility to the table. Traditional shaders are more rigid in their application, usually tailored for rendering operations. Compute shaders, however, can perform more generalized computations. This is crucial in SPH, where complex operations for density, pressure, and force calculations are commonplace.

Another noteworthy benefit is the performance gain achieved through hardware acceleration. The architecture of modern GPUs is designed to handle a high degree of parallelism, significantly speeding up the overall computation time for the SPH simulation.

The general approach to implementing SPH using compute shaders begins with the initialization of data buffers to store particle information, such as positions, velocities, and densities. A dedicated compute shader then initializes these particle states. Following the initialization, a sequence of compute shaders are dispatched in the pipeline for tasks such as neighbor finding, density calculation, intermediate velocity calculation, and pressure gradient determination. Each of these shaders updates the particle states stored in the buffer. Finally, the particle positions and velocities are updated in a concluding shader pass.

For data storage, Shader Storage Buffer Objects (SSBOs) offer read-write access to large datasets, making them suitable for storing particle data. For constant data, such as physical constants or simulation parameters, Uniform Buffer Objects (UBOs) are more appropriate.

Synchronization is a key consideration in this context. It's crucial to ensure that all writes to shared data structures are properly synchronized to avoid race conditions.

In summary, the compute shader offers a highly flexible and performant medium for SPH implementation, capable of dealing with the algorithm's complexity and computational intensity. Through the thoughtful application of compute shaders, significant performance gains can be realized, propelling the SPH simulation closer to real-time capabilities.

V. Results

The outcomes of implementing Smoothed Particle Hydrodynamics (SPH) using various computational platforms are quite illuminating. This section provides a comprehensive look at the performance comparison between compute shaders, multi-threaded CPU, and single-threaded CPU implementations. Specifically, We evaluated the time taken to update the system state for 75,000 particles in each of these computational paradigms. (Fig. 1)

In terms of timing, compute shaders have shown a staggering performance advantage. The GPU-based compute shader implementation completes an update in approximately 10 milliseconds. In stark contrast, the multi-threaded CPU implementation takes about 250 milliseconds for the same number of particles, despite the parallelization advantages offered by multi-threading. Even more strikingly, the single-threaded CPU implementation lags far behind, taking approximately 690 milliseconds to complete an update. (Fig.1)

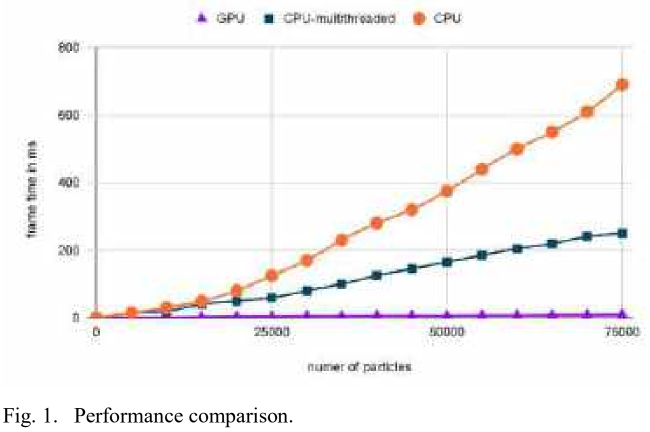


Fig. 1. Performance comparison.

The results clearly indicate the superiority of using compute shaders for SPH in terms of computational time. The speedup is not just incremental; it's rather transformative, bringing the simulation much closer to real-time capabilities. It is worth mentioning that the hardware acceleration facilitated by the architecture of modern GPUs plays a crucial role here. The inherent parallelism in SPH is exploited to its fullest extent by the compute shaders, which is something even multi-threaded CPU implementations find hard to match.

Conclusion

The overarching objective of this article was to scrutinize the computational performance of SPH algorithms under different execution frameworks. Compute shaders emerged as the unequivocal winner in terms of computational speed, showcasing the remarkable capabilities of GPU computing for such complex, particle-based simulations.

By implementing SPH using compute shaders, multi-threaded CPUs, and single-threaded CPUs, the study unveiled the striking time efficiencies gained with the use of compute shaders. Specifically, compute shaders managed to update 75,000 particle states within 10 milliseconds, thereby drastically reducing computational time compared to the other approaches.

The findings clearly underline the advantage of adopting modern GPU-based methodologies for real-time simulations, especially in fields that require quick data throughput. This is not merely an incremental progression but signifies a momentous stride in SPH computations.

Interestingly, this heightened efficiency did not compromise the simulation's fidelity. The quality of the results remained consistent across the board, suggesting that the benefits of using compute shaders go beyond just speed. The findings also pave the way for future research to delve into other domains where speed and real-time processing are paramount.

To sum up, the research validates the utility of compute shaders in the efficient and accurate execution of SPH algorithms. This opens new avenues not only for SPH but also for other computational models that can capitalize on the immense processing power of modern GPUs. As we move towards an era where computational demand is ever-increasing, the insights from this study serve as a crucial benchmark and an inspiration for future endeavors in high-performance computing.

References

[1] M. S. Bartlett. “Statistical estimation of density functions”. In: Sankhya: The Indian Journal of Statistics, Series A (1963), pp. 245–254

[2] W. Hu, G. Guo, X. Hu, D. Negrut, (2019). A Consistent, Spatially Adaptive Smoothed Particle Hydrodynamics Method for Fluid-Structure Interactions. Computer Methods in Applied Mechanics and Engineering. 347. 402-424. 10.1016/j.cma.2018.10.049.

[3] J. Monaghan, (2005). Smoothed Particle Hydrodynamics. Reports on Progress in Physics. 68. 1703. 10.1088/0034-4885/68/8/R01.

[4] S. Gunadi, P. Yugopuspito, (2018). Real-Time GPU-based SPH Fluid Simulation Using Vulkan and OpenGL Compute Shaders. 1-6. 10.1109/ICSTC.2018.8528699.

[5] R. Akhunov, R. Winchenbach, A. Kolb, (2023). Evaluation of particle-based smoothed particle hydrodynamics boundary handling approaches in computer animation. Computer Animation and Virtual Worlds. 10.1002/cav.2138.

[6] R. A. Gingold, J. J. Monaghan, “Smoothed particle hydrodynamics: theory and application to non-spherical stars,” In: Monthly Notices of the Royal Astronomical Society 181.3 (1977), pp. 375–389 (Cited on pages 2, 31, 50, 70, 71, 88, 89, 117).

Machine Learning as One of the Highly Effective Methods of Reducing the Load on CSOC's Analysts

Roman Karpiuk
Department of cybersecurity
Ivan Franko National University of Lviv
Lviv, Ukraine
simmppllee@gmail.com

Petro Venherskyi
Department of cybersecurity
Ivan Franko National University of Lviv
Lviv, Ukraine
petro,vengersky@gmail.com

Anna Korchenko
Department of information security and telecommunications
National Technical University Dnipro Polytechnic
Dnipro, Ukraine
annakor@ukr.net

Yuliia Khokhlachova
Department of information technology security
National Aviation University
Kyiv, Ukraine
yuliiahohlachova@gmail.com

Abstract — **The primary objective of this investigation revolves around streamlining the operational workflow within the Cybersecurity Operation Center (CSOC). It is no secret that the CSOC faces a significant challenge due to the influx of signals originating from a multitude of cybersecurity tools, each demanding precise processing. These tools encompass Intrusion Detection Systems (IDS), Endpoint Detection and Response (EDR), Next-Generation Firewalls (NGFW), Data Loss Prevention (DLP), Cloud Access Security Brokers (CASB), and more. Furthermore, a substantial volume of raw information, including event logs from diverse systems and applications, necessitates analysis and decision-making. This cumulative workload places immense pressure on CSOC analysts, resulting in an upsurge in poorly processed events, longer response times, extended event processing durations, and inevitably, an increase in the number of false positives. To address these challenges, two viable options emerge:**

1. **Augment CSOC funding by recruiting additional analysts. However, this approach is not without its hurdles, including a scarcity of qualified specialists in the job market and the potential for inflated financial costs, which may not align with optimal business decisions.**
2. **Develop an integrated system designed to detect malicious actions comprehensively. A key component of such a system involves the sophisticated detection of anomalies and responding not solely to individual events but to the anomalies themselves.**

Keywords* — *cybersecurity, framework, machine learning (ML), Security Information and Event Management (SIEM), Splunk, Cybersecurity Operation Center (CSOC).

I. Introduction

How can one discern the optimal approach for detecting the activities of adversarial teams? What is the ideal number of security operations (SecOps) tools to deploy? And, most crucially, how can an organization ensure it possesses the technical and human resources necessary to effectively handle the influx of events generated by these tools? The singular solution lies in the establishment of a comprehensive cybersecurity program within the organization. This program should be underpinned by well-defined policies and procedures, delineated staff responsibilities, and a diverse range of technological frameworks. However, what course of action should be taken when all these elements are in place, yet the organization's internal systems expand, its workforce grows, the array of tools employed multiplies, and consequently, the volume of data requiring analysis by cybersecurity analysts surges?

II. Problem Statement

When an organization experiences growth across all quantitative metrics, spanning from the proliferation of servers to an expanding fleet of end-user machines, attackers interpret this as a singular opportunity. It signifies that they can maneuver through an extensive array of techniques encompassed within each tactical segment defined by MITRE ATT&CK.

For cybersecurity professionals, this reality translates to an escalation in occurrences, particularly false positives, with each passing day. The task of distinguishing genuine threats from the cacophony of "noise" becomes progressively more challenging. Consequently, organizations face two primary strategies to address this predicament on a global scale:

1. Increase the size of their Cybersecurity Operations Center (CSOC) workforce.
2. Optimize the workloads of existing CSOC analysts.

Given that not every organization can allocate additional resources to their cybersecurity program, we will focus on the latter approach.

The foremost consideration in alleviating the burden on analysts is the reduction of alerts generated by various security operations (SecOps) tools. However, organizations often encounter a significant challenge when they position themselves as "enterprises" and opt for cybersecurity solutions that are predominantly commercial. These solutions typically lack flexibility in influencing the alert volume generated by individual systems. They function as enigmatic "black boxes," and it would be ill-advised to attempt alterations to their logic, given the substantial engineering teams behind their development.

The remaining avenue involves working with the end result produced by these systems, specifically, the alerts. Most threat detection engineers traverse several stages of refinement to curtail the alert count, including:

a) Adjusting threshold values.

b) Calculating various statistical metrics such as average (avg), minimum (min), maximum (max), standard deviation (stdev), and others.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

c) Employing statistical comparisons across different timeframes.

d) Implementing machine learning.

In this context, our focus will be directed towards effectively harnessing machine learning techniques to identify anomalies, both within raw event logs and amidst extensive clusters of alerts.

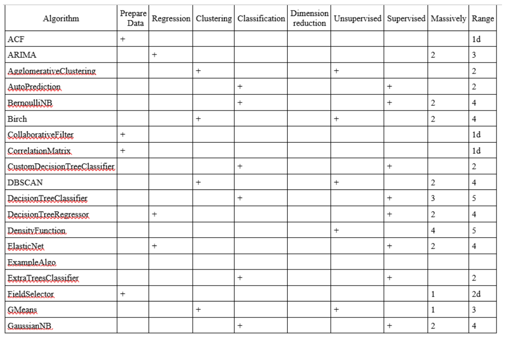
III. Prepare Your Paper Before Styling

The cornerstone of building an effective cyber-anomaly detection system lies in data. To detect anomalies with a probability greater than 40%, it's essential to gather consistent data spanning a duration of at least 90 days. Extending the data collection timeframe beyond 90 days is even more advantageous. However, it's crucial to acknowledge that amassing a substantial volume of data can pose challenges in terms of computational power and machine learning calculations, particularly lengthening the training time. Therefore, the optimal choice would be a 180-day timeframe.

A 180-day period encompasses numerous dynamic changes within the IT infrastructure, including password changes, maintenance activities, periodic alterations, and, inevitably, atypical actions. Subsequently, the next step involves constructing statistics that facilitate the identification of desired anomalies. These statistics must be highly accurate and free from anticipated noise, such as typical anomaly actions like anomalous failed logins by certain systems or applications. This meticulous approach to data collection and analysis enhances the precision of cyber-anomaly detection, enabling organizations to focus on genuine threats while minimizing false positives..

Once the data has been prepared, and statistics have been generated from the dataset, the next critical step is to determine which machine learning (ML) algorithm is best suited for anomaly detection. While there are over 80 machine learning algorithms available, not all of them are suitable for identifying anomalies. Some ML algorithms excel in prediction or forecasting tasks, while others are more adept at clustering.

To make an informed choice, it's essential to categorize these algorithms into classifications based on their suitability for anomaly detection. This categorization can help streamline the selection process and ensure that the chosen algorithm aligns with the specific goals of anomaly detection within your dataset. By narrowing down the options and selecting algorithms optimized for anomaly detection, you can improve the accuracy and effectiveness of your cybersecurity system.



Algorithm	Prepare Data	Regression	Clustering	Classification	Dimension reduction	Unsupervised	Supervised	Massively	Range
ACF	+								1d
ARIMA		+						2	3
AgglomerativeClustering			+			+			2
AutoPrediction				+			+		2
BernoulliNB				+			+	2	4
Birch			+			+		2	4
CollaborativeFilter	+								1d
CorrelationMatrix	+								1d
CustomDecisionTreeClassifier				+			+		2
DBSCAN			+			+		2	4
DecisionTreeClassifier				+			+	3	5
DecisionTreeRegressor		+					+	2	4
DensityFunction						+		4	5
ElasticNet		+					+	2	4
ExampleAlgo									
ExtraTreesClassifier				+			+		2
FieldSelector	+							1	2d
GMeans			+			+		1	3
GaussianNB				+			+	2	4

Fig. 1. Classification of Machine Learning Algorithms in the Cybersecurity Domain

Following the classification process, the careful elimination of unsuitable algorithms, and rigorous testing using real-world data, it has become evident that one of the most effective algorithms for detecting anomalies is the density function.

Given its demonstrated effectiveness and reliability in our context, we have chosen to employ the density function as a core component of our framework. This choice aligns with our goal of developing a robust and accurate anomaly detection system within the cybersecurity domain. The density function will serve as a pivotal tool in our efforts to identify and address potential threats and anomalies effectively.

Our framework, consisting of six steps, provides a systematic approach to effective anomaly detection within the cybersecurity domain:

1. Dataset Preparation and General Statistics: in this initial step, you gather and prepare the dataset, ensuring its quality and consistency. You then build general statistics to gain insights into the data.

2. Data Analysis and Noise Reduction:following dataset preparation, you analyze the dataset to identify and reduce noise, ensuring that the data is clean and relevant for further processing.

3. Feature Extraction: in this phase, you extract useful features (UF) from the data, which are essential for the subsequent stages of the anomaly detection process.

4. Dataset Generalization: here, you generalize the dataset based on the identified objects and the extracted useful features. This step helps streamline the data for efficient analysis.

5. Algorithm Training: this critical step involves training the machine learning algorithm with the prepared and generalized data, ensuring that it can recognize patterns and anomalies during the designated timeslot.

6. Application of Pre-trained ML Model: in the final step, you apply the pre-trained machine learning model to the data that requires processing. This model, equipped with the knowledge gained during training, can effectively detect anomalies within the dataset.

Framework provides a structured approach to cybersecurity anomaly detection, emphasizing data quality, feature extraction, and the use of machine learning to enhance threat identification and response.

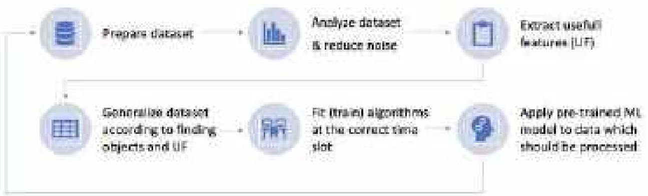

Prepare dataset
Analyze dataset & reduce noise
Extract usefull features (UF)
Generalize dataset according to finding objects and UF
Fit (train) algorithms at the correct time slot
Apply pre-trained ML model to data which should be processed

Fig. 2. Anomaly detection with machine learning.

The initial two steps have been previously detailed.

Moving on to the third step, it revolves around calculating essential parameters that enhance the granularity and robustness of our statistical analysis for the observed object. Essentially, this step involves extracting and incorporating additional fields into our dataset and subsequent stages. These additional attributes serve to dissect our statistics into finer segments, providing a more accurate and selective view. For instance, we can calculate whether events occur during the day or night, at the beginning or end of the week, on weekdays or weekends, and break them down by hours or specific days of the week. These supplementary attributes empower us to scrutinize our statistical sample with greater precision, enabling us to pinpoint anomalies with greater accuracy.

The fourth step is indispensable for generalizing our statistics across various dimensions, including objects, useful features (UF), and time. In the first two steps, we construct datasets based on unique IDS events, where "unique" is determined by a combination of source, destination, and signature. This particular phase generalizes the dataset across attributes like <source> if our aim is to uncover anomalies in a unique source. Alternatively, it generalizes the dataset across <signature> if we are seeking anomalies triggered by specific signatures during specific time patterns, such as weekends. The flexibility here allows us to adapt our approach to various scenarios based on data processing requirements and our specific objectives.

The fifth step is the fitting stage, which constitutes the primary process for training our algorithm using the finalized dataset. Timing is the linchpin of this step. It's crucial to precisely define when we expect to encounter anomalies. This is because the construction of statistics greatly varies between, for example, a one-hour time interval and a 24-hour one. The distinction becomes even more pronounced when detecting anomalies across weeks or months (as in the case of TX, where we seek to identify attacker actions that correspond to intelligent evasion tactics). Importantly, it's imperative not to formulate a training sample around one time interval and then attempt to identify anomalies using another time interval. Such an approach would lead to fundamentally flawed operations.

The sixth and final step involves the application of a pre-trained model for direct anomaly detection. In this phase, we must complete the initial four steps as well. Another critical consideration is that the training data should exclude instances where anomalies have been detected. In essence, the training sample, for instance, is constructed over a 180-day period, omitting data from the present day (<today> minus one day), while the trained model operates on the data for the current day.

This structured approach ensures the effectiveness of our cybersecurity anomaly detection framework, emphasizing the significance of data preparation, feature engineering, training, and model application in addressing evolving security threats accurately.

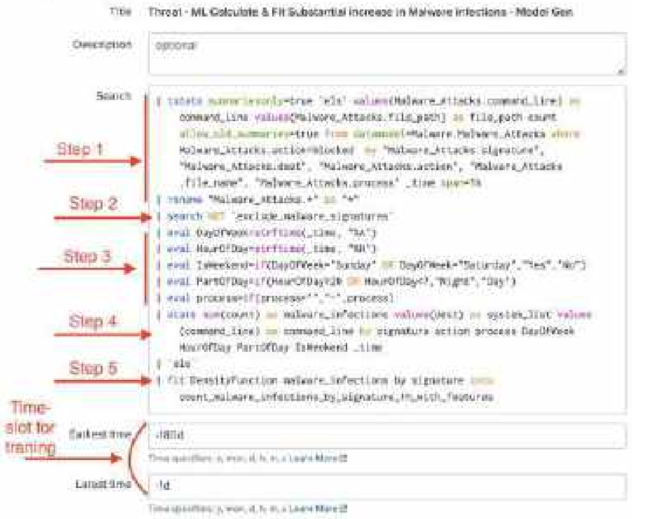
IV. Real Examples and Testing

A. *Anomaly Detection Across Unique EDR <signature>*

Testing Environment:

- Splunk Enterprise Security (a SIEM application by Splunk) [2]
- Machine Learning Toolkit [2]
- Raw data sourced from the Organization's Endpoint Detection and Response System (EDR).

In these scenarios, we leverage data from our EDR system that has been carefully normalized to conform to Splunk's data model for "Malware" [2]. This approach grants us the advantage of working with standardized fields as defined by Splunk. As a result, our search processes are expedited, and we no longer rely on EDR vendors for data compatibility.



Title: Threat - ML Calculate & Fit Substantial Increase in Malware Infections - Model Gen

Description: optional

Search:

```
| tstats summariesonly=true `els` values(Malware_Attacks.command_line) as
    command_line values(Malware_Attacks.file_path) as file_path count
    allow_old_summaries=true from datamodel=Malware.Malware_Attacks where
    Malware_Attacks.action=blocked by "Malware_Attacks.signature",
    "Malware_Attacks.dest", "Malware_Attacks.action", "Malware_Attacks
    .file_name", "Malware_Attacks.process" _time span=1h
| rename "Malware_Attacks.*" as "*"
| search NOT `exclude_malware_signatures`
| eval DayOfWeek=strftime(_time, "%A")
| eval HourOfDay=strftime(_time, "%H")
| eval IsWeekend=if(DayOfWeek="Sunday" OR DayOfWeek="Saturday","Yes","No")
| eval PartOfDay=if(HourOfDay>20 OR HourOfDay<7,"Night","Day")
| eval process=if(process="","-",process)
| stats sum(count) as malware_infections values(dest) as system_list values
    (command_line) as command_line by signature action process DayOfWeek
    HourOfDay PartOfDay IsWeekend _time
| `els`
| fit DensityFunction malware_infections by signature into
    count_malware_infections_by_signature_1h_with_features
```

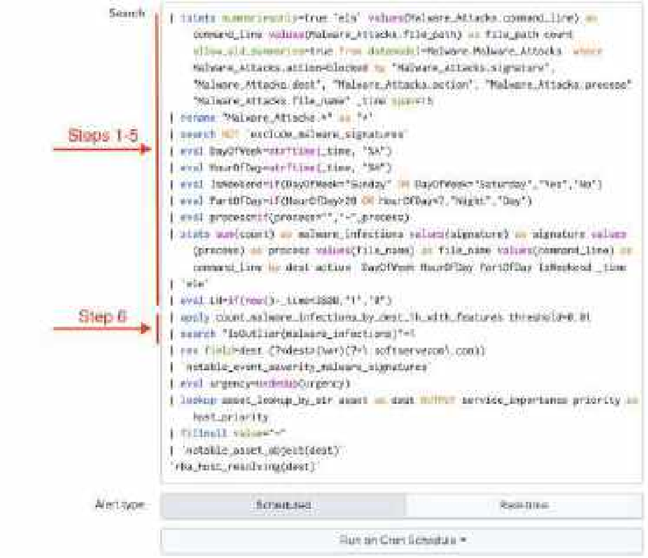
Earliest time: -180d

Time specifiers: y, mon, d, h, m, s Learn More

Latest time: -1d

Time specifiers: y, mon, d, h, m, s Learn More

Fig. 3. Splunk correlation search for training ML algorithm utilizing EDR data from standardized data model



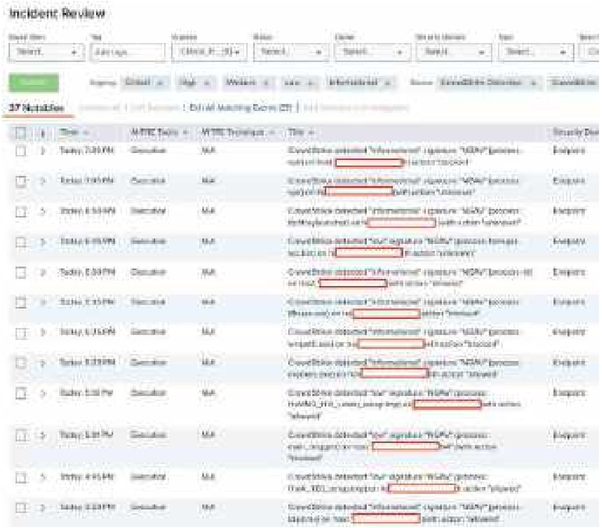
Search:

```
| tstats summariesonly=true `els` values(Malware_Attacks.command_line) as
    command_line values(Malware_Attacks.file_path) as file_path count
    allow_old_summaries=true from datamodel=Malware.Malware_Attacks where
    Malware_Attacks.action=blocked by "Malware_Attacks.signature",
    "Malware_Attacks.dest", "Malware_Attacks.action", "Malware_Attacks.process"
    "Malware_Attacks.file_name" _time span=1h
| rename "Malware_Attacks.*" as "*"
| search NOT `exclude_malware_signatures`
| eval DayOfWeek=strftime(_time, "%A")
| eval HourOfDay=strftime(_time, "%H")
| eval IsWeekend=if(DayOfWeek="Sunday" OR DayOfWeek="Saturday","Yes","No")
| eval PartOfDay=if(HourOfDay>20 OR HourOfDay<7,"Night","Day")
| eval process=if(process="","-",process)
| stats sum(count) as malware_infections values(signature) as signature values
    (process) as process values(file_name) as file_name values(command_line) as
    command_line by dest action DayOfWeek HourOfDay PartOfDay IsWeekend _time
| `els`
| eval id=if(now()-_time<3600,"1","0")
| apply count_malware_infections_by_dest_1h_with_features threshold=0.01
| search "IsOutlier(malware_infections)"=1
| rex field=dest (?<dest>(\w+)(?=\.softserveinc\.com))
| `notable_event_severity_malware_signatures`
| eval urgency=lower(urgency)
| lookup asset_lookup_by_str asset as dest OUTPUT service_importance priority as
    host_priority
| fillnull value="-"
| `notable_asset_object(dest)`
`rba_host_resolving(dest)`
```

Alert type: Scheduled | Real-time

Run on Cron Schedule ▾

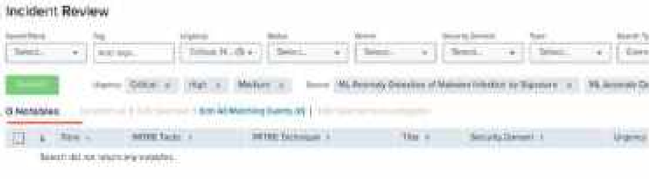
Fig. 4. Splunk correlation search for malware anomaly detection using pre-trained ML model



Incident Review

37 Notables

Fig. 5. Standard EDR reaction. 37 events for last 24 hours which should be processed by Tier1 analysts (before ML framework was implemented)



Incident Review

0 Notables

Search did not return any notables.

Fig. 6. Result of anomaly detection. 0 event for last 24 hours which should be processed by Tier1 analysts (after ML framework implementation)

As evident from our analysis, it's not necessary to individually investigate each EDR detection, often considered as noise. Instead, we can employ a scoring mechanism, such as the Risk-Based Approach (RBA), to assign scores to this noise. By doing so, our focus can shift towards detecting anomalies and findings with high RBA scores through correlation searches driven by machine learning.

B. *Anomaly Detection Across Kerberos and LDAP Requests*

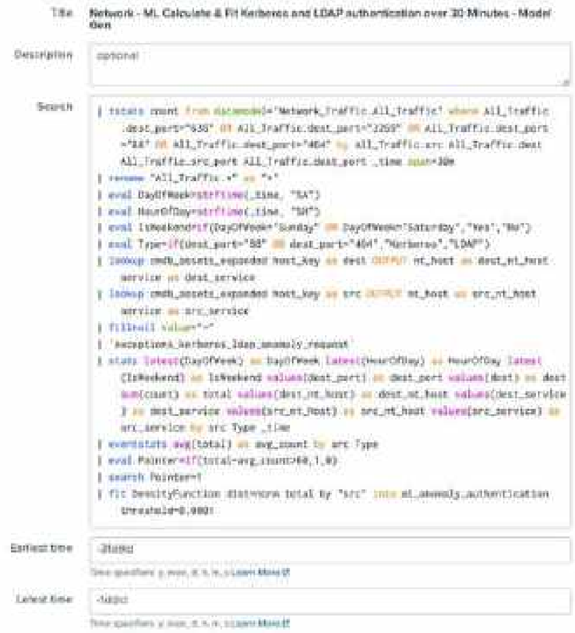
Testing Environment:

- Splunk Enterprise Security [2]
- Machine Learning Toolkit [2]
- Organizational logs retrieved from the Next-Generation Firewall (NGFW).

In a parallel manner to the previous scenario, our primary objective remains the construction of highly robust statistics within our dataset. Our aim is to eliminate noise and maintain a comprehensive understanding of our objectives and the data under scrutiny.

We retain the flexibility to amalgamate a multitude of parameters within our detection mechanism to ensure adaptability to our infrastructure or specific requirements.

Subsequent to training the machine learning model, which has been transformed into <ml_anomaly_authentication>, we commence the anomaly detection process. This process entails generating notable events that are subsequently passed on for further investigation by Tier 1 CSOC analysts.



Title: Network - ML Calculate & Fit Kerberos and LDAP authentication over 30 Minutes - Model Gen

Description: optional

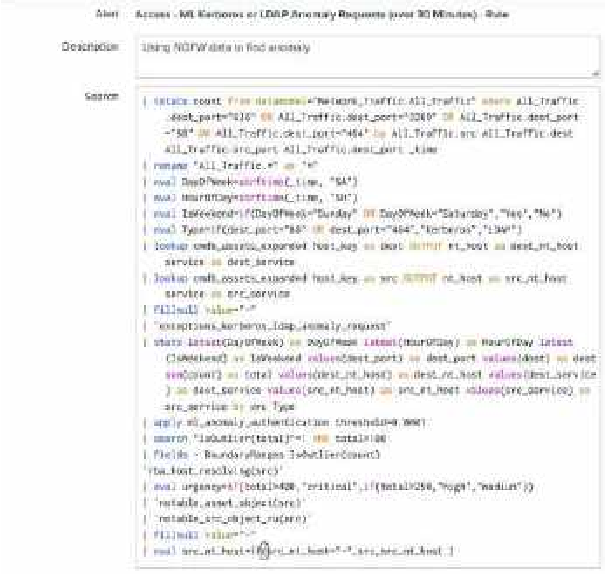
Search:

```
| tstats count from datamodel="Network_Traffic.All_Traffic" where All_Traffic
    .dest_port="636" OR All_Traffic.dest_port="3269" OR All_Traffic.dest_port
    ="88" OR All_Traffic.dest_port="464" by All_Traffic.src All_Traffic.dest
    All_Traffic.src_port All_Traffic.dest_port _time span=30m
| rename "All_Traffic.*" as "*"
| eval DayOfWeek=strftime(_time, "%A")
| eval HourOfDay=strftime(_time, "%H")
| eval IsWeekend=if(DayOfWeek="Sunday" OR DayOfWeek="Saturday","Yes","No")
| eval Type=if(dest_port="88" OR dest_port="464","Kerberos","LDAP")
| lookup cmdb_assets_expanded host_key as dest OUTPUT nt_host as dest_nt_host
    service as dest_service
| lookup cmdb_assets_expanded host_key as src OUTPUT nt_host as src_nt_host
    service as src_service
| fillnull value="-"
| `exceptions_kerberos_ldap_anomaly_request`
| stats latest(DayOfWeek) as DayOfWeek latest(HourOfDay) as HourOfDay latest
    (IsWeekend) as IsWeekend values(dest_port) as dest_port values(dest) as dest
    sum(count) as total values(dest_nt_host) as dest_nt_host values(dest_service
    ) as dest_service values(src_nt_host) as src_nt_host values(src_service) as
    src_service by src Type _time
| eventstats avg(total) as avg_count by src Type
| eval Pointer=if(total-avg_count>60,1,0)
| search Pointer=1
| fit DensityFunction dist=norm total by "src" into ml_anomaly_authentication
    threshold=0.001
```

Earliest time: -31d@d

Latest time: -1d@d

Fig. 7. Splunk correlation search for training ML algorithm utilizing NGFW data from standardized data model



Alert: Access - ML Kerberos or LDAP Anomaly Requests (over 30 Minutes) - Rule

Description: Using NGFW data to find anomaly

Search:

```
| tstats count from datamodel="Network_Traffic.All_Traffic" where All_Traffic
    .dest_port="636" OR All_Traffic.dest_port="3269" OR All_Traffic.dest_port
    ="88" OR All_Traffic.dest_port="464" by All_Traffic.src All_Traffic.dest
    All_Traffic.src_port All_Traffic.dest_port _time
| rename "All_Traffic.*" as "*"
| eval DayOfWeek=strftime(_time, "%A")
| eval HourOfDay=strftime(_time, "%H")
| eval IsWeekend=if(DayOfWeek="Sunday" OR DayOfWeek="Saturday","Yes","No")
| eval Type=if(dest_port="88" OR dest_port="464","Kerberos","LDAP")
| lookup cmdb_assets_expanded host_key as dest OUTPUT nt_host as dest_nt_host
    service as dest_service
| lookup cmdb_assets_expanded host_key as src OUTPUT nt_host as src_nt_host
    service as src_service
| fillnull value="-"
| `exceptions_kerberos_ldap_anomaly_request`
| stats latest(DayOfWeek) as DayOfWeek latest(HourOfDay) as HourOfDay latest
    (IsWeekend) as IsWeekend values(dest_port) as dest_port values(dest) as dest
    sum(count) as total values(dest_nt_host) as dest_nt_host values(dest_service
    ) as dest_service values(src_nt_host) as src_nt_host values(src_service) as
    src_service by src Type
| apply ml_anomaly_authentication threshold=0.0001
| search 'IsOutlier(total)'=1 AND total>100
| fields - BoundaryRanges IsOutlierCount)
`rba_host_resolving(src)`
| eval urgency=if(total>400,"critical",if(total>250,"high","medium"))
| `notable_asset_object(src)`
| `notable_src_object_ru(src)`
| fillnull value="-"
| eval src_nt_host=if(src_nt_host="-",src,src_nt_host )
```

Fig. 8. Splunk correlation search for Kerberos/LDAP anomaly detection using pre-trained ML model

C. *TimeToTriage, TimeToClosure and False-Positive Ratios Improvements*

The ultimate validation of machine learning's efficacy lies in the reduction of the workload imposed on the CSOC. Specifically, this entails a decrease in the number of triggered correlation searches, particularly in cases where ML-based rules are applied. Furthermore, when ML-driven correlation rules do activate, the false positive rate remains below 50%, resulting in an overall reduction in false alarms. This, in turn, enables feasible investments in TTT (Time to Triage) and TTC (Time to Contain) strategies, with intervals as short as 15 and 30 minutes, respectively—adhering to industry best practices.

It's important to emphasize that the performance metrics and descriptions presented here are derived from real organizational data within a specific company and its unique IT infrastructure, distinguishing it from open datasets like those provided by MITRE or others.

However, in a broader context, the conclusion remains unequivocally positive. The adoption of correlation rule optimization practices, where applicable, invariably enhances the detection of potentially malicious activities, without necessitating substantial additional financial investments. It's a swift victory that benefits both the CSOC and the organization as a whole.

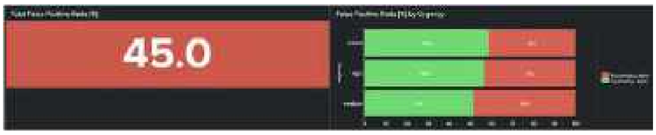


Fig. 9. False-Positive ratio for last 90 days

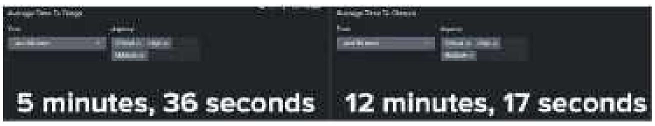


Fig. 10. TTT&TTC results for last 90 days

REFERENCES

[1] S. Haider, S. Ozdemir, Hands-On Machine Learning for Cybersecurity. Packt Publishing Ltd, 2018.

[2] “Make your organization more resilient”, [Online]. Available: https://www.splunk.com

[3] “Machine Learning for Cybersecurity”, [Online]. Available: https://towardsdatascience.com/machine-learning-for-cybersecurity-101-7822b802790b

[4] “Machine Learning: Practical Applications for Cybersecurity”, [Online]. Available: https://www.recordedfuture.com/machine-learning-cybersecurity-applications/

Risks' Attribute Values Evaluation in Software Engineering by Monte Carlo Simulation

Maria Lyashkevych
System Design Department
Ivan Franko Lviv National University
Lviv, Ukraine
https://orcid.org/0000-0002-9655-036X

Vasyl Lyashkevych
System Design Department
Ivan Franko Lviv National University
Lviv, Ukraine
https://orcid.org/0000-0003-2810-6061

Roman Shuvar
System Design Department
Ivan Franko Lviv National University
Lviv, Ukraine
https://orcid.org/0000-0001-6768-4695

Abstract **— It is a good practice in the industry to carry out a risk assessment even before the stages of software production because we can prevent the costs of both human and material resources. In the era of the success of artificial intelligence with machine learning, it is very important to have accumulated expert knowledge on the basis of which we fuel datasets. Usually, it is very difficult to obtain numerical values of risk attributes, and especially to determine their importance among others. The values of risk attributes can be balanced in various ways, one of which is Monte Carlo simulation, which is presented in the publication.**

Keywords — risk assessment, software engineering risks, risk calculation, Monte Carlo simulation, value assessments, risk modelling

I. Introduction

IT software development projects face risks due to the complexity of the project, the lack of a full understanding of the laws of the global development environment, long code development, and the propensity of scientists to complex methodologies. Science and engineering are the main drivers of these applications, with developers often not funded or trained in software development. [1]

In fact, many projects begin as research projects, without a clear relationship between deliverables, schedule, and resources. Developers are also often users or a large part of the user community. The development process is largely driven by prototypes, which are constantly refined and refined by scientists.

Software development is a complex process that involves a multitude of interrelated factors, making it one of the most intricate and challenging fields of engineering. [2] Thus, in early development phases, design complexity metrics is considered useful indicators of software testing effort and some quality attributes. [3] Software complexity is an important factor which ought to be recognized at different levels of software development. [4]

II. Risks in Software Development

A. Common Risks in Software Development

Software development is a complex and multifaceted process, and it involves various risks that can impact the success of a project. These risks can manifest at different stages of development and can be technical, organizational, or external in nature. [5] The most widespread common risks in software development are below:

- Unclear Requirements. Poorly defined or constantly changing requirements can lead to misunderstandings, scope creep, project delays, and budget overruns.
- Inadequate Planning. Lack of thorough project planning can result in underestimated budgets, unrealistic schedules, and resource shortages.
- Scope Creep. Uncontrolled expansion of project scope can lead to missed deadlines, increased costs, and a project that never seems to be completed.
- Technical Complexity. Complex technical requirements, unfamiliar technologies, or dependencies on third-party components can pose significant challenges and risks to the project's success.
- Resource Constraints. Insufficient or poorly allocated resources (including skilled personnel, hardware, and software) can lead to project delays and subpar results.
- Inadequate Testing. Skipping or insufficient testing can result in the release of software with critical defects, leading to customer dissatisfaction and costly post-release fixes.
- Security Vulnerabilities. Failure to address security risks can lead to data breaches, system vulnerabilities, and compromised user information.
- Communication Issues. Poor communication among team members and stakeholders can lead to misunderstandings, missed requirements, and project delays.
- Dependency Risks. Reliance on external libraries, APIs, or third-party services can introduce risks if those dependencies change or become unavailable.
- Personnel Turnover. Frequent turnover of team members can disrupt project continuity and knowledge transfer, potentially leading to project delays and quality issues.
- Technology Obsolescence. Rapid advancements in technology can lead to the obsolescence of tools and technologies used in the project, necessitating costly rework.
- Legal and Compliance Risks. Non-compliance with legal and regulatory requirements can result in legal issues, fines, and project delays.
- Budget Overruns. Poor cost estimation and financial management can lead to budget overruns and jeopardize the financial health of the project.
- Market Changes. Changes in market conditions, customer preferences, or competition can render the software less relevant or require significant adjustments.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

- Natural Disasters and External Events. Events like natural disasters, economic crises, or global pandemics can disrupt project schedules and resource availability.
- Poor Vendor or Partner Performance. If you rely on third-party vendors or partners for critical components or services, their underperformance can have a cascading effect on your project.
- Stakeholder Expectations. Misalignment of stakeholder expectations can lead to dissatisfaction with the final product, even if it technically meets the requirements.

Risks can be classified as systematic and unsystematic risks. Systematic risks involve external factors like hacking, viruses, natural disasters, and power loss, while unsystematic risks involve unique risks such as misuse of confidential data, application errors, inside attacks, data loss, equipment malfunctions, and human interactions. The main goal of risk management is to identify and control all possible risks before they occur during software development.

B. *Reasons for the Complexity Factors*

To manage these risks effectively, software development teams often employ risk assessment and mitigation strategies. This may include conducting thorough risk assessments at the project's outset, implementing robust change control processes, fostering strong communication and collaboration, and continuously monitoring and adapting to changing circumstances throughout the project's lifecycle. Therefore, when assessing the risks in software development, we should pay attention to the complexity factors which could be described by the following reasons:

- Software inherently deals with abstract concepts and logical structures, which can be inherently complex to design, implement, and maintain. Unlike physical systems, software doesn't have the same tangible components, making it difficult to visualize and understand. [6]
- Software projects typically involve diverse stakeholders with varying needs and expectations. These stakeholders may include end-users, customers, managers, developers, testers, designers, and more. Managing and aligning their interests can be complex.
- Software requirements often change over time due to evolving business needs, user feedback, or market dynamics. Adapting to these changes while maintaining project stability is a significant challenge. [7]
- The rapid pace of technological advancements necessitates continuous learning and adaptation. Developers must stay up-to-date with new programming languages, frameworks, libraries, and tools.
- In many software projects, integrating various software components, modules, or third-party services can be challenging due to compatibility issues, data synchronization, and dependencies. [8]
- Comprehensive testing is critical to ensure software quality, but it can be complex due to various factors, including the need for diverse testing approaches (unit, integration, system, etc.), test data generation, and handling edge cases. [2, 8]
- As software systems grow, scalability and performance optimization become complex issues. Ensuring that the software can handle increased loads and maintain responsiveness is a constant concern.
- Protecting software from security threats, including vulnerabilities, cyberattacks, and data breaches, is an ongoing challenge that requires expertise and vigilance.
- Managing software projects involves coordinating tasks, resources, timelines, and budgets. Ensuring that projects are delivered on time and within budget can be complex due to changing requirements and uncertainties. [9]
- Proper documentation and knowledge transfer are essential for maintaining and enhancing software over time. Keeping documentation up-to-date and ensuring knowledge continuity across team members is often overlooked but critical.
- Dealing with legacy systems, which may lack proper documentation or use outdated technologies, can be complex when upgrading or integrating them with modern solutions. [10]
- Some software projects, especially in industries like healthcare or finance, must adhere to strict regulatory compliance standards, adding complexity in terms of documentation, validation, and auditing.
- Designing software that provides an intuitive and pleasing user experience requires a deep understanding of user needs and preferences, adding a layer of complexity to the development process.
- Building software for global markets may involve adapting it to different languages, cultures, and legal requirements, which adds complexity to the development process.
- Successful software development often requires a blend of technical and non-technical skills, including project management, communication, problem-solving, and creativity. [9]

Indeed, software development is complex due to its abstract concepts, diverse stakeholders, evolving business needs, rapid technological advancements, and integrating components. Managing projects, maintaining documentation, dealing with legacy systems, adhering to regulatory compliance standards, and designing intuitive user experiences are all challenges. Adapting to global markets and balancing technical and non-technical skills is essential for successful software development.

III. Widespread Risk Identification Approaches

A. *Taxonomy-based Risk Identification*

The taxonomy-based risk identification was a popular approach to risk assessment in software engineering because the taxonomy of software development maps the characteristics of software development and hence of software development risks. [6-11] The risk identification method is based on the following assumptions [11]:

- Software development risks are generally known by the project’s technical staff but are poorly communicated.
- A structured and repeatable method of risk identification is necessary for consistent risk management.
- Effective risk identification must cover all key development and support areas of the project.
- The risk identification process must create and sustain a non-judgmental and non-attributive risk elicitation environment so that tentative or controversial views are heard.
- No overall judgment can be made about the success or failure of a project based solely on the number or nature of risks uncovered.

This report [11] describes a method for facilitating the systematic and repeatable identification of risks associated with the development of a software-dependent project. Results of the field tests encouraged the claim that the described method is useful, usable, and efficient.

Choosing the right software technology stack and managing your project are crucially important things to thoroughly understand the project’s requirements, including functionality, scalability, security, performance, and any specific constraints or preferences, and engage stakeholders. [12] Ultimately, the right technology stack will depend on the unique requirements and goals of your project. It's essential to strike a balance between meeting current needs and preparing for future growth and changes. Regularly reassess your technology choices as your project evolves to ensure they remain appropriate.

Software requirements can significantly impact the development process, quality, and project success. Common impacts include project delays, budget overruns, reduced quality, scope creep, customer dissatisfaction, legal and regulatory issues, resource allocation issues, communication breakdown, testing challenges, technical debt, project abandonment, and reputation damage. Unclear or changing requirements can lead to rework, increased development costs, reduced quality, scope creep, and customer dissatisfaction. Non-compliance with legal or regulatory requirements can result in fines, lawsuits, or costly rework. Incomplete or changing requirements can also lead to suboptimal resource allocation, communication breakdowns, testing challenges, technical debt, project abandonment, and reputation damage. To mitigate these impacts, effective requirements engineering and management practices, such as involving stakeholders early, using transparent documentation, conducting thorough analysis, and implementing change control processes, are crucial. [13]

B. Risks Modelling

Modelling risks in software development is a crucial step in the risk management process. [13] By creating risk models in software development, you can identify potential risks, categorize them, create a risk register, assess their probability and impact, calculate risk exposure, prioritize risks, model dependencies, develop mitigation strategies, create contingency plans, assign risk owners, monitor the risk register, communicate with stakeholders, model risk impact, conduct sensitivity analyses, use risk management tools, schedule regular risk reviews, and conduct a post-project review. [14-15] This comprehensive risk model helps manage and mitigate risks throughout the project, increasing the likelihood of successful delivery. Developing a comprehensive risk model to manage and mitigate software development project risks, we should revise and refine them iteratively evolving new risks arise.

No one is immune to risk, and businesses with poor risk management are not protected. Common software development risks include pitfalls and bottlenecks in each industry. Generally, they are:

- Planning. Bad timing in software development can lead to significant profits or setbacks. To avoid this, use agile methodologies, involve team members, receive feedback, and involve stakeholders.
- Budget estimation. Improper budget estimation can lead to project completion late or exceeding the agreed cost. To mitigate this, maintain control, discuss additional functions, and calculate costs at the discussion stage.
- Professional skills. Poor code quality and technical risks in software development can lead to negative consequences, including lack of professionalism, constant changes in software requirements, inadequate development support, complex projects, and difficult implementation. Poor productivity in software development projects can be caused by poor project management, incorrect methodology, and mismatched team members. Agile methodologies can maintain motivation and productivity, while project managers can mentor and coach the team.
- Engagement and management. Poor project management leads to 32% of project failures, resulting in employee turnover and delays. High standards for project managers include strategic and tactical skills, strong communication, organizational framework, and documentation. User engagement is crucial for software development success but can be risky due to insufficient research, incorrect solution selection, or outdated UX/UI design. Professional business analysts analyse target audience needs and usability.
- Unexpectability. Unpredictable external risks in software development include market changes, competitor growth, government regulations, and consumer behaviour. Business analysts use technologies like Machine Learning and Big Data Analytics to analyse market trends and make informed decisions. A well-thought-out risk management strategy can significantly reduce project impact. [15]

C. Machine Learning approach

Machine learning plays a crucial role in risk assessment for software development by utilizing historical data, patterns, and predictive models to identify and manage potential risks. It can predict the likelihood of specific risks occurring, identify them using natural language processing (NLP) and text mining techniques, categorize risks into predefined categories, assess their severity, estimate time-to-impact, optimize resource allocation, rank and prioritize risks, build early warning systems, provide real-time risk reporting, detect anomalous behaviours or patterns in project data, recommend optimized risk mitigation strategies, continuously improve risk management practices, offer data-driven decision making, and scale to handle large volumes of project data. [16]

Machine learning models can also help in identifying emerging risks or deviations from the norm by analyzing project data and providing insights into risk assessment and mitigation. This allows for informed decisions based on objective data and helps in balancing workloads and managing critical resources effectively.

In addition to risk prediction, machine learning can also help in resource allocation by considering predicted risks and their impact on project tasks. It can also help in identifying anomalous behaviours or patterns in project data, leading to more effective risk management plans.

To leverage machine learning effectively for risk assessment in software development, it's essential to have access to high-quality historical data and to invest in data preprocessing, model training, and ongoing model evaluation and refinement. Additionally, collaboration between domain experts and data scientists is crucial to ensure that machine learning models align with the specific needs and context of software development projects.

IV. The Datasets Availability

Risk is a significant issue that can lead to significant losses and threats in various organizational procedures, particularly in the Computer Science field. Risks can arise from networks, the internet, malicious codes, users, loopholes, and physical security. High-quality software systems can be created with various risks, but project managers can reduce their impact by calculating these risks on IT resources. The software industry is one of the world's largest industries which sells an average of $350 billion of off-the-shelf software annually. [17-18]

Risk management is crucial for large-scale systems due to their high quality and reliability demands. Software project assessment and prediction systems often rely on past project analysis results to form formulae using statistical techniques. However, credibility is often given to larger datasets, and little consideration is given to adding new project results or removing them from a dataset. Gathering the dataset, we should analyse the construction and use of historical software project data repositories in case study companies, using provided guidelines on the formation and usage of these datasets. [19]

The risk analysis techniques quantify the quality of these datasets, discussing the expected reliability of results and how this can be used to formulate dataset policies. We should aim to provide a comprehensive understanding of the importance of historical software project data repositories in software project assessment and prediction systems [20].

Software risk management is a crucial practice in the software industry that involves risk identification, estimation, mitigation, and monitoring. It provides a disciplined environment for efficient decision-making in software development. Large-scale systems are particularly challenging due to the complexity of risks and the different risk factors they have. This paper provides an exhaustive list of risk factors for large-scale and small-scale systems and presents a comparative analysis of different software-related risk management models. The models are categorized based on the severity of their risks, highlighting the importance of understanding and addressing these risks in software development. [21]

The paper [22] examines the construction and use of historical software project data repositories in various case study companies. It aims to provide guidelines on the formation of such repositories and the appropriateness of adding new project results or removing project results from a dataset. The authors argue that while large datasets are often considered more credible, the authors also highlight the importance of considering when to add or remove project results from a dataset. The paper [22] concludes that a more comprehensive approach to risk and software metrics datasets is needed to improve the credibility of these systems.

A data set for risk assessment in software development is a collection of historical data and information that can be used to analyse and predict risks associated with software projects. These datasets often include data on project attributes such as size, complexity, team size, development methodology, and outcome variables related to project success or failure. Some common types of datasets used for risk assessment in software development include defect datasets, project outcome datasets, effort estimation datasets, requirements and change request datasets, personnel datasets, process and methodology datasets, code metrics datasets, vendor or supplier datasets, external factors datasets, and security datasets. Finally, we gathered the accessible datasets through the Internet and after the feature analysis picked up the best of them in a single dataset for our targets.

V. Monte Carlo Simulation

Monte Carlo simulation is a powerful technique for modelling and analyzing complex systems by using random numbers. It's widely used in various fields such as finance, physics, engineering, and computer science. Monte Carlo simulations are a class of computational techniques that use random sampling to approximate and analyse complex systems or problems.

While Monte Carlo simulations themselves are not typically considered machine learning algorithms, they can be combined with machine learning techniques to enhance their capabilities. It's important to note that the specific choice of machine learning algorithms in Monte Carlo simulations depends on the nature of the problem, the available data, and the objectives of the simulation. The risk attributes could be assessed in different ways using categorical, nominal or another type of value representation.

Of course, we would like to have a numerical value for each of the risk's attributes because it is easy to do calculations about them and apply powerful algorithms such as machine learning algorithms. We can guess that using the Monte Carlo simulations we can investigate the best possible types of values within the risk assessment models. As we have a lot of different criteria we used different data distributions for the initial random generation of the attribute values, for example normal, exponential, Poisson, binomial, geometric, gamma, beta, lognormal and triangular.

Different algorithms may be more suitable for different scenarios (Fig. 1), and their selection should be based on careful analysis and experimentation.

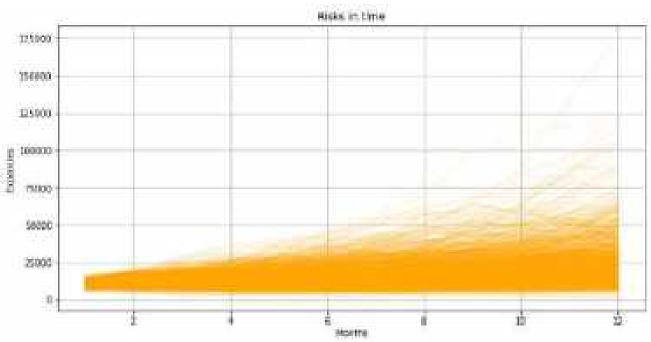


Fig. 1. Results of Monte Carlo simulation

As can be seen in Fig. 1, we inspected the values of possible expenses which we could get with applied risks based on the generated initial values. Changing the distributions and specific limitations we found the best approach for risk attribute values representation.

Conclusions

Applying the Monte Carlo simulation, we resolved some pain items of our problem. We clarified the attributes of software engineering risks choosing the best data distribution and investigated diapasons for attribute values.

Achieved results allow us to investigate the risk attribute importance and create a good dataset for applying the machine learning algorithms.

References

[1] Kemerer, C.F. Software complexity and software maintenance: A survey of empirical research. Ann Software Eng 1, 1–22 (1995). https://doi.org/10.1007/BF02249043

[2] Liu Enfei. Risk Factors of Software Development Projects in Chinese IT Small and Medium-Sized Enterprises. KTH Royal Institute of Technology, 2015. 54 p.

[3] Anh Nguyen-Duc. The impact of software complexity on cost and quality – a comparative analysis between open source and proprietary software. International Journal of Software Engineering & Applications (IJSEA), vol.8, no.2, March 2017. pp.17-31. https://doi.org/10.5121/ijsea.2017.8202

[4] Uk, Ijeacs. (2016). Software Complexity Measurement: A Critical Review. International Journal of Engineering and Applied Computer Science (IJEACS). 01. 12-16.

[5] M. Bhasi. A study on software development project risk, risk management, project outcomes and their inter-relationship, 2010, 46 p.

[6] Jain, Rashmi & Dey, Sujoy. (2004). A Life-Cycle Taxonomy for Assessing Software Development Risks.

[7] Menezes Júnior, Júlio & Gusmao, Cristine & Moura, Hermano. (2019). Risk factors in software development projects: a systematic literature review. Software Quality Journal. 27. https://doi.org/10.1007/s11219-018-9427-5

[8] Maniasi, Sebastian & Britos, Paola & Garcia-Martinez, Ramon. (2006). A Taxonomy-Based Model for Identifying Risks. 13-18.

[9] Luís M. Alves, Gustavo Souza, Pedro Ribeiro, Ricardo J. Machado. Longevity of risks in software development projects: a comparative analysis with an academic environment, Procedia Computer Science, Volume 181, 2021, pp. 827-834. https://doi.org/10.1016/j.procs.2021.01.236

[10] Richard P. Kendall. A Proposed Taxonomy for Software Development Risks for High-Performance Computing (HPC) Scientific/Engineering Applications. Carnegie Mellon, 2007, 39 p.

[11] Marvin J. Carr Suresh L. Konda Ira Monarch F. Carol Ulrich. Taxonomy-Based Risk Identification. Technical Repor, 1993, 90 p.

[12] O.V. Pomorova, A.V. Ivanov, M.Yu. Lyashkevych. Fuzzy inference system for analysis, monitoring and risk assessment in software development // Bulletin of Khmelnytskyi National University. Technical Sciences, 2013, no 1, pp. 93-100.

[13] Sanson, Miguel. "ISO 31010 2019 Risk Management -Risk Assessment Techniques Management Du Risque -Techniques d'Appréciation Du Risque." ISO 31010 2019 GESTIÓN DEL RIESGO TÉCNICAS DE EVALUACIÓN (2019): n. pag. Print.

[14] Saad Yasser Chadli and Ali Idri. 2017. Identifying and mitigating risks of software project management in global software development. In Proceedings of the 27th International Workshop on Software Measurement and 12th International Conference on Software Process and Product Measurement (IWSM Mensura '17). Association for Computing Machinery, New York, NY, USA, 12–22. https://doi.org/10.1145/3143434.3143453

[15] Jafari, Sajad. Analysis of Risk Factors in Global Software Development: A Cross-Continental Study Using Modified Firefly Algorithm. Computational Intelligence and Neuroscience, 2022. https://doi.org/10.1155/2022/4936748

[16] Mahmud, M.H.; Nayan, M.T.H.; Ashir, D.M.N.A.; Kabir, M.A. Software Risk Prediction: Systematic Literature Review on Machine Learning Techniques. Appl. Sci. 2022, 12, 11694. https://doi.org/10.3390/app122211694

[17] Breno Gontijo Tavares, Mark Keil, Carlos Eduardo Sanches da Silva & Adler Diniz de Souza (2021) A Risk Management Tool for Agile Software Development, Journal of Computer Information Systems, 61:6, 561-570, DOI: 10.1080/08874417.2020.1839813

[18] Sandra Miranda Neves, Carlos Eduardo Sanches da Silva. Risk management applied to software development projects in incubated technology-based companies: literature review, classification, and analysis. Gest. Prod., São Carlos, vol. 23, no. 4, Pp. 798-814, 2016 http://dx.doi.org/10.1590/0104-530X472-15

[19] M. Pasha, G. Qaiser and U. Pasha, "A Critical Analysis of Software Risk Management Techniques in Large Scale Systems," in IEEE Access, vol. 6, pp. 12412-12424, 2018, https://doi.org/10.1109/ACCESS.2018.2805862

[20] Shaukat, Zain & Naseem, Rashid & Khan, Muhammad Zubair. (2018). A Dataset for Software Requirements Risk Prediction. 112-118. DOI: 10.1109/CSE.2018.00022.

[21] D.G. Edgar-Nevill. Risk And Software Metrics Datasets. WIT Transactions on Information and Communication Technologies, vol. 8, 1994. https://doi.org/10.2495/SQM940231

[22] Zain Shaukat, Rashid Naseem, & Muhammad Zubair. (2018). Software Requirement Risk Prediction Dataset (Version 1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.1209601

Gaussian Mixture Model Based Machine Learning Approach for Detection of Threat Types in Communication Networks

Oleksii Holubnychyi
Telecommunication and Radioelectronic Systems Department
National Aviation University
Kyiv, Ukraine
oleksii.holubnychyi@npp.nau.edu.ua

Maksym Zaliskyi
Telecommunication and Radioelectronic Systems Department
National Aviation University
Kyiv, Ukraine
maximus2812@ukr.net

Oleksandr Solomentsev
Telecommunication and Radioelectronic Systems Department
National Aviation University
Kyiv, Ukraine
avsolomentsev@ukr.net

Ivan Ostroumov
Air Navigation Systems Department
National Aviation University
Kyiv, Ukraine
ostroumovv@ukr.net

Yuliya Averyanova
Air Navigation Systems Department
National Aviation University
Kyiv, Ukraine
ayua@nau.edu.ua

Olha Sushchenko
Aerospace Control Systems Department
National Aviation University
Kyiv, Ukraine
sushoa@ukr.net

***Abstract* — The Gaussian mixture model (GMM) based machine learning approach for automatic unsupervised detection of threat types in communication networks is proposed in the paper. The proposed approach uses the expectation-maximization algorithm and a vector of threat associated parameters, which are monitored parameters in communication network. These parameters can be associated with both features of the functioning of communication network and threat scenarios, e.g., delays and number of requests in cases of either specific routing or threats. The detection of threat types in the proposed approach is performed by a forming of subsets of threat associated parameters using latent variables of GMM, which are by nature posterior probabilities that threat associated parameters are associated with components of the GMM. An example of simulation of the proposed approach, which deals with a detection of possible "no threats", "insignificant threat", "moderate threat", and "significant threat" cases in communication network is shown and analyzed. Features and prospects of the proposed approach are also shown and analyzed in the paper.**

Keywords —cybersecurity, machine learning, threat detection, Gaussian mixture model, expectation-maximization algorithm

I. Introduction

Modern communication networks are made out as groups of communication techniques and layers (physical, data link, network, transport, session etc.).

Different techniques in communication networks can be associated with different specific kinds of technical issues such as spectrum sharing in cognitive radio networks [1], designing UAV-assisted 5G hybrid data link structures [2], improvement of radio resource usage in hybrid wireless data link structures [3], data analysis [4], [5] in communication [6], navigation [7], [8] and surveillance [9], [10] air traffic management (CNS/ATM) systems, sensor networks [11] etc.

Abovementioned techniques foresee various layers of networks, including layers of heterogenous [1]-[3], ISO/OSI, TCP/IP structures [12], [13]. This, for example, leads to a high complexity of cybersecurity features in modern 5G communications [14] and airborne systems where machine learning (ML) and artificial intelligence (AI) techniques, in particular neural networks, become more prevalent [15]-[17].

II. Problem Statement

Threats for privacy and security cybersecurity in modern communication networks can have various causes and nature, starting with the first layer of the ISO/OSI model, which is responsible for signal transferring and processing, and ending with the seventh layer of the ISO/OSI model, which is responsible for user applications and relevant interconnections. This causes a multivariate approach to cybersecurity problem, which take into account complex analysis of different threat types and their automatic unsupervised estimation and detection. Such complex analysis can be realized directly by means of ML [18] and AI [19] techniques, where only observed input data and, in some instances, pre-configuration of ML and AI techniques are required.

The maximum likelihood estimation (MLE) approach is often used as the basis for a number of ML techniques, in particular for an unsupervised regression analysis and clustering [20].

In doing so, a clustering of threat types within the multivariate approach to cybersecurity problem can be based on MLE approach [21], [22]. This area is a promising direction for an automatic unsupervised detection of threat types in communication networks.

The problem boils down to the justification and adjustment of MLE based ML technique for detection and further clustering of threat types in communication networks.

In this regard, the aim of the paper is to propose an adjusted MLE based ML approach for detection and further clustering of threat types in communication networks.

A proposed approach can be based on a Gaussian mixture model (GMM) [23], [24] along with the expectation-maximization (EM) algorithm [22], [25], [26].

III. Methodology

Let $\mathbf{TAP} = (TAP_1, TAP_2, \dots, TAP_N)^{\mathrm{T}}$ be a vector of N threat associated parameters TAP_n, $n = 1,2,\dots,N$, which are monitored parameters in communication network.

These parameters can be associated with both features of the functioning of communication network and threat

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

scenarios, e.g., delays and number of requests in cases of either specific routing or threats (e.g., denial-of-service attacks).

The vector $\mathbf{TAP}$ can contain normalized values TAP_n, $n = 1,2,\dots,N$, in some metric, where these parameters can be presented and analyzed.

A kind of normed metric depends on a structure and features both communication network and expandable threats, e.g., the Hankel matrix rank can be used as a metric for the Hankel-based unsupervised anomaly detection [27].

The proposed MLE based ML approach primarily automatically forms subsets of threat associated parameters within $\mathbf{TAP}$. This may make it possible to further detect various specific subsets of threat associated parameters, which characterize relevant threat types.

The forming of subsets of threat associated parameters within $\mathbf{TAP}$ using a GMM [23], [24] is realized through representing the data in $\mathbf{TAP}$ as a mixture of Gaussian distributions:

$$PDF(TAP) = \sum_{k=1}^{K} \frac{W_k}{\sqrt{2\pi D_k}} \exp\left[-\frac{(TAP-M_k)^2}{2D_k}\right], \qquad (1)$$

where $PDF(TAP)$ is the probability density function of mixture of Gaussian distributions for the random variable TAP; K is the number of GMM components; M_k and D_k are mean and variance of k-th GMM component, respectively; W_k is the weighting coefficient of k-th GMM component, and these coefficients meet the condition $\sum_{k=1}^{K} W_k = 1$.

As with most statistical data processing models [28]-[31], GMM is often portraited as a statistical model, which can be used for the parameter estimation.

However, there are algorithms, which work with GMM and could make it possible to determine belongings of threat associated parameters TAP_n, $n = 1,2,\dots,N$, to GMM components.

In particular, EM algorithm [22], [25], [26], in addition to parameter estimation of M_k, D_k, and W_k, $k = 1,2,\dots,K$, in (1), can also give estimates of GMM latent variables.

The latent variables are by nature posterior probabilities that TAP_n is associated with k-th GMM component. This feature can be used for detection and further clustering of threat types in communication networks.

In this context, each GMM component plays a role of a threat type.

EM algorithm maximizes the log-likelihood function $LLF(\mathbf{M},\mathbf{D},\mathbf{W}|\mathbf{TAP})$ for a conjecture for GMM parameters $\mathbf{M},\mathbf{D}$ and $\mathbf{W}$ at observed threat associated parameters $\mathbf{TAP}$:

$$\begin{aligned} LLF(\mathbf{M},\mathbf{D},\mathbf{W}|\mathbf{TAP}) &= \log \prod_{n=1}^{N} PDF(TAP_n|\mathbf{M},\mathbf{D},\mathbf{W}) = \\ &= \sum_{n=1}^{N} \log \sum_{k=1}^{K} \frac{W_k}{\sqrt{2\pi D_k}} \exp\left[-\frac{(TAP-M_k)^2}{2D_k}\right] \to \max. \end{aligned} \qquad (2)$$

Finding the maximum of $LLF(\mathbf{M},\mathbf{D},\mathbf{W}|\mathbf{TAP})$ is done through GMM latent variables $\mathbf{\Theta} = (\theta_{n,k})$, $n = 1,2,\dots,N$, $k = 1,2,\dots,K$, by means of two consecutive iterative procedures, namely expectation step (E-step) and maximization step (M-step) [22, 25].

- E-step:

$$\theta_{n,k}^{(q-1)} = \frac{\frac{W_k^{(q-1)}}{\sqrt{2\pi D_k^{(q-1)}}} \exp\left[-\frac{\left(TAP_n - M_k^{(q-1)}\right)^2}{2D_k^{(q-1)}}\right]}{\sum_{m=1}^{K} \frac{W_m^{(q-1)}}{\sqrt{2\pi D_m^{(q-1)}}} \exp\left[-\frac{\left(TAP_n - M_m^{(q-1)}\right)^2}{2D_m^{(q-1)}}\right]}, \qquad (3)$$

where q is the iteration number, $q \in \mathbb{N}$.

Initial parameters $\mathbf{M}^{(0)} = \left(M_k^{(0)}\right)$, $\mathbf{D}^{(0)} = \left(D_k^{(0)}\right)$, and $\mathbf{W}^{(0)} = \left(W_k^{(0)}\right)$, $k = 1,2,\dots,K$, in (3) can be chosen using a priori data about the most probable or expected their values, taking into account statistics of possible threat types and their parameters in a communication network.

The number of GMM components K is an internal parameter of a cybersecurity analysis system, which can be taken for a number of threat types. For instance, for the case $K = 1$ a cybersecurity analysis system in communication network does not detect different threat types, and, for the case $K = N$, a cybersecurity analysis system is able to detect each threat associated parameter as a separate threat.

The detection and further clustering of threat types in communication networks allow organizing of protection for different threat types when subsets of threat associated parameters can be automatically identified as threat types within $\mathbf{TAP}$, e.g., at $K = 4$: "no threats" ($k = 1$), "insignificant threat" ($k = 2$), "moderate threat" ($k = 3$), and "significant threat" ($k = 4$).

The estimated at the q-th iteration number of threat associated parameters, which belong to the k-th GMM component and relevant k-th threat type, is expressed as:

$$Q_k^{(q-1)} = \sum_{n=1}^{N} \theta_{n,k}^{(q-1)}. \qquad (4)$$

- M-step:

$$W_k^{(q)} = Q_k^{(q-1)}/N; \qquad (5)$$

$$M_k^{(q)} = \frac{1}{Q_k^{(q-1)}} \sum_{n=1}^{N} \theta_{n,k}^{(q-1)} TAP_n; \qquad (6)$$

$$D_k^{(q)} = \frac{1}{Q_k^{(q-1)}} \sum_{n=1}^{N} \theta_{n,k}^{(q-1)} \left(TAP_n - M_k^{(q)}\right)^2 \qquad (7)$$

The E-step and M-step are recurrent and periodically repeated until the stopping criteria (8) is not met.

$$\begin{aligned} &LLF\left(\mathbf{M}^{(q)},\mathbf{D}^{(q)},\mathbf{W}^{(q)}\middle|\mathbf{TAP}\right) - \\ &-LLF\left(\mathbf{M}^{(q-1)},\mathbf{D}^{(q-1)},\mathbf{W}^{(q-1)}\middle|\mathbf{TAP}\right) < \xi, \end{aligned} \qquad (8)$$

where ξ is a positive small number under which the convergence of the EM algorithm can be considered acceptable.

IV. Simulation Results and Their Analysis

A. Description of the Example

Let us simulate the proposed MLE based ML approach for detection and further clustering of threat types in communication network.

Suppose that the observed in communication network **TAP** contains $N = 24$ measured threat associated parameters, which are, e.g., numbers of server requests per unit time:

$$\mathbf{TAP} = (13, 3, 25, 35, 28, 917, 15, 840, 14, 17, 985, 11, \\ 89, 31, 60, 16, 45, 51, 783, 51, 7, 59, 53, 21)^{\mathrm{T}}. \tag{9}$$

If using the mentioned above cybersecurity analysis system, which is characterized by the internal parameter $K = 4$ with the threat types conditional classification containing "no threats" case ($k = 1$), "insignificant threat" case ($k = 2$), "moderate threat" case ($k = 3$), and "significant threat" case ($k = 4$), the initial parameters $\mathbf{M}^{(0)}$, $\mathbf{D}^{(0)}$, and $\mathbf{W}^{(0)}$ are primarily subject for justification and setting. This is also the main component of adjustment of the proposed ML approach for an automatic and unsupervised detection and further clustering of threat types in communication networks.

Let, for example, typical a priori expected values of $\mathbf{M}^{(0)}$, which are taken into account with the statistics of possible threat types and their parameters in a communication network, are presented in Table I. It is relevant to note that values of $\mathbf{M}^{(0)}$ can vary immensely, but they must not be the same in order to avoid degeneration of different GMM components, when some or all of them become the same.

The initial approximation for variances $\mathbf{D}^{(0)}$ can be taken as $D_k^{(0)} \sim \left[\max\left(M_1^{(0)}, M_2^{(0)}, \dots, M_K^{(0)}\right)\right]^2$, $k = 1, 2, \dots, K$, i.e., for the considered example $D_k^{(0)} = \left[M_1^{(0)}\right]^2 = 10^6$, $k = 1, 2, 3, 4$. Such large values of $\mathbf{D}^{(0)}$ in fact transform normally distributed initial GMM components into uniformly distributed components under conditions of a priori uncertainty about **TAP**. This provides a good initial grip of each **TAP** component by each GMM component.

The initial approximation for weighting coefficients under conditions of a priori uncertainty about **TAP** can be taken as $W_k^{(0)} = 1/K$, $k = 1, 2, \dots, K$, i.e., these coefficients are evenly distributed, and for the considered example $W_k^{(0)} = 1/4$, $k = 1, 2, 3, 4$.

TABLE I. EXAMPLE OF INITIAL PARAMETERS $\mathbf{M}^{(0)}$

GMM component	Threat type	$\boldsymbol{M}^{(0)}$
$k = 1$	No threats	1
$k = 2$	Insignificant threat	10
$k = 3$	Moderate threat	100
$k = 4$	Significant threat	1000

B. Simulation Results

The following results are presented for the case $\xi = 0.001$.

$LLF\left(\mathbf{M}^{(0)}, \mathbf{D}^{(0)}, \mathbf{W}^{(0)} \middle| \mathbf{TAP}\right) = -190.8625.$

- Results after the iteration $q = 1$:

$\mathbf{Q}^{(0)} = (6.395, 6.403, 6.460, 4.741)^{\mathrm{T}};$

$\mathbf{W}^{(1)} = (0.266, 0.267, 0.269, 0.198)^{\mathrm{T}};$

$\mathbf{M}^{(1)} = (148, 149, 157, 264)^{\mathrm{T}};$

$\mathbf{D}^{(1)} = (8.62 \cdot 10^4, 8.67 \cdot 10^4, 9.16 \cdot 10^4, 1.46 \cdot 10^5)^{\mathrm{T}};$

$LLF\left(\mathbf{M}^{(1)}, \mathbf{D}^{(1)}, \mathbf{W}^{(1)} \middle| \mathbf{TAP}\right) = -171.6985.$

Stopping criteria (8) is not fulfilled.

- Results after the iteration $q = 2$:

$\mathbf{Q}^{(1)} = (6.329, 6.330, 6.327, 5.015)^{\mathrm{T}};$

$\mathbf{W}^{(2)} = (0.264, 0.264, 0.264, 0.209)^{\mathrm{T}};$

$\mathbf{M}^{(2)} = (110, 112, 131, 387)^{\mathrm{T}};$

$\mathbf{D}^{(2)} = (6.00 \cdot 10^4, 6.12 \cdot 10^4, 7.39 \cdot 10^4, 1.82 \cdot 10^5)^{\mathrm{T}};$

$LLF\left(\mathbf{M}^{(2)}, \mathbf{D}^{(2)}, \mathbf{W}^{(2)} \middle| \mathbf{TAP}\right) = -169.0545.$

Stopping criteria (8) is not fulfilled.

- Results after the iteration $q = 3$:

$\mathbf{Q}^{(2)} = (6.342, 6.286, 5.839, 5.533)^{\mathrm{T}};$

$\mathbf{W}^{(3)} = (0.264, 0.262, 0.243, 0.231)^{\mathrm{T}};$

$\mathbf{M}^{(3)} = (48, 50, 77, 561)^{\mathrm{T}};$

$\mathbf{D}^{(3)} = (1.29 \cdot 10^4, 1.44 \cdot 10^4, 3.51 \cdot 10^4, 1.76 \cdot 10^5)^{\mathrm{T}};$

$LLF\left(\mathbf{M}^{(3)}, \mathbf{D}^{(3)}, \mathbf{W}^{(3)} \middle| \mathbf{TAP}\right) = -156.0108.$

Stopping criteria (8) is not fulfilled.

In the following, results for 3 latest iterations, when the convergence of EM algorithm is achieved, are shown.

- Results after the iteration $q = 78$:

$\mathbf{Q}^{(77)} = (8.734, 2.417, 8.849, 4.000)^{\mathrm{T}};$

$\mathbf{W}^{(78)} = (0.364, 0.101, 0.369, 0.167)^{\mathrm{T}};$

$\mathbf{M}^{(78)} = (13, 29, 52, 881)^{\mathrm{T}};$

$\mathbf{D}^{(78)} = (28.90, 12.70, 306.95, 5849.19)^{\mathrm{T}};$

$LLF\left(\mathbf{M}^{(78)}, \mathbf{D}^{(78)}, \mathbf{W}^{(78)} \middle| \mathbf{TAP}\right) = -120.0318.$

Stopping criteria (8) is not fulfilled.

- Results after the iteration $q = 79$:

$\mathbf{Q}^{(78)} = (8.758, 2.399, 8.843, 4.000)^{\mathrm{T}};$

$\mathbf{W}^{(79)} = (0.365, 0.100,0.368, 0.167)^{\mathrm{T}};$

$\mathbf{M}^{(79)} = (13, 29, 52, 881)^{\mathrm{T}};$

$\mathbf{D}^{(79)} = (29.10, 12.48, 306.46, 5849.19)^{\mathrm{T}};$

$LLF\left(\mathbf{M}^{(79)}, \mathbf{D}^{(79)}, \mathbf{W}^{(79)} \middle| \mathbf{TAP}\right) = -120.0306.$

Stopping criteria (8) is not fulfilled.

- Results after the iteration $q = 80$:

$\mathbf{Q}^{(79)} = (8.777, 2.384, 8.839, 4.000)^{\mathrm{T}};$

$\mathbf{W}^{(80)} = (0.366, 0.099, 0.368, 0.167)^{\mathrm{T}};$

$\mathbf{M}^{(80)} = (13, 29, 52, 881)^{\mathrm{T}};$

$\mathbf{D}^{(80)} = (29.27, 12.32, 306.08, 5849.19)^{\mathrm{T}};$

$LLF\left(\mathbf{M}^{(80)}, \mathbf{D}^{(80)}, \mathbf{W}^{(80)} \middle| \mathbf{TAP}\right) = -120.0298.$

Stopping criteria (8) is fulfilled.

Obtained values of GMM latent variables are:

$$\mathbf{\Theta}^{(79)}=\begin{bmatrix}0.975 & 0.000 & 0.025 & 0.000\\ 0.965 & 0.000 & 0.035 & 0.000\\ 0.215 & 0.559 & 0.226 & 0.000\\ 0.001 & 0.315 & 0.684 & 0.000\\ 0.041 & 0.738 & 0.221 & 0.000\\ 0.000 & 0.000 & 0.000 & 1.000\\ 0.966 & 0.000 & 0.034 & 0.000\\ 0.000 & 0.000 & 0.000 & 1.000\\ 0.971 & 0.000 & 0.029 & 0.000\\ 0.947 & 0.002 & 0.052 & 0.000\\ 0.000 & 0.000 & 0.000 & 1.000\\ 0.979 & 0.000 & 0.021 & 0.000\\ 0.000 & 0.000 & 1.000 & 0.000\\ 0.008 & 0.688 & 0.303 & 0.000\\ 0.000 & 0.000 & 1.000 & 0.000\\ 0.958 & 0.001 & 0.041 & 0.000\\ 0.000 & 0.000 & 1.000 & 0.000\\ 0.000 & 0.000 & 1.000 & 0.000\\ 0.000 & 0.000 & 0.000 & 1.000\\ 0.000 & 0.000 & 1.000 & 0.000\\ 0.978 & 0.000 & 0.022 & 0.000\\ 0.000 & 0.000 & 1.000 & 0.000\\ 0.000 & 0.000 & 1.000 & 0.000\\ 0.774 & 0.081 & 0.144 & 0.000\end{bmatrix}. \qquad (10)$$

Detection and clustering of threat types in communication networks can be implemented by means of detection of belongings of threat associated parameters TAP_n, $n = 1,2,\dots,N$, to GMM components, i.e., to subsets of threat associated parameters within $\mathbf{TAP}$. For this purpose, for each n-th threat associated parameter as the criterion of maximum posterior probability can be applied the criterion $\max(\theta_{n,1}, \theta_{n,2}, \theta_{n,3}, \theta_{n,4})$. Analysis of $\mathbf{\Theta}^{(79)}$ in (10) using this criterion gives results of threat types detection, which are shown in Table II.

C. Analysiss of Simulation Results and Discussion

Results in Table II mean that:

- "No threats" is characterized by values of threat associated parameters in the range [3; 21];
- "Insignificant threat" is characterized by values of threat associated parameters in the range [25; 31];
- "Moderate threat" is characterized by values of threat associated parameters in the range [35; 89];
- "Significant threat" is characterized by values of threat associated parameters in the range [783; 985].

TABLE II. RESULTS OF THREAT TYPE DETECTION

GMM component	Threat type	Threat associated parameters
$k = 1$	No threats	$TAP_1 = 13$; $TAP_2 = 3$; $TAP_7 = 15$; $TAP_9 = 14$; $TAP_{10} = 17$; $TAP_{12} = 11$; $TAP_{16} = 16$; $TAP_{21} = 7$; $TAP_{24} = 21$
$k = 2$	Insignificant threat	$TAP_3 = 25$; $TAP_5 = 28$; $TAP_{14} = 31$
$k = 3$	Moderate threat	$TAP_4 = 35$; $TAP_{13} = 89$; $TAP_{15} = 60$; $TAP_{17} = 45$; $TAP_{18} = 51$; $TAP_{20} = 51$; $TAP_{22} = 59$; $TAP_{23} = 53$
$k = 4$	Significant threat	$TAP_6 = 917$; $TAP_8 = 840$; $TAP_{11} = 985$; $TAP_{19} = 783$

Attention should be paid to the main distinctive feature of the proposed GMM MLE based ML approach, which is automatic unsupervised organizing of subsets of threat associated parameters within $\mathbf{TAP}$ that are corresponding to threat types in communication networks (i.e., their detection in this way).

This organizing is also implemented without any previously setting of thresholds between ranges, and using, in fact, only typical a priori expected values of threat associated parameters for each threat type.

Another important feature of the proposed ML approach is the ability to estimate the accuracy of threat types detection. For this purpose, it is possible to use for each k-th threat type the following criteria:

- $\mathrm{MAX}_k = \max(\theta_{n,k} | TAP_n \in \text{GMM component } k)$;
- $\mathrm{MIN}_k = \min(\theta_{n,k} | TAP_n \in \text{GMM component } k)$;
- $\mathrm{MEAN}_k = \langle \theta_{n,k} | TAP_n \in \text{GMM component } k \rangle$.

Results of estimation the accuracy of threat types detection for the considered example is presented in Table III.

In the case of using the proposed criteria, higher values of MAX_k, MIN_k, or MEAN_k indicate higher accuracy of threat types detection.

The proposed ML approach is also characterized by the following features:

- convergence rate of iterative procedures within EM algorithm in the proposed ML approach is acceptable: few dozen of iterations for detection and further clustering of all threat types in communication networks are required; e.g., analysis of $\mathbf{TAP}$, which consists of $N = 24$ observed threat associated parameters, took no more than a hundred iterations; however, the number of iterations significantly depends on observed data $\mathbf{TAP}$ and initial parameters $\mathbf{M}^{(0)}$, $\mathbf{D}^{(0)}$, $\mathbf{W}^{(0)}$; this number is increasing for a larger number of N and larger number of threat types;
- forming of subsets of threat associated parameters within $\mathbf{TAP}$ can be a robust and accurate: a decision about belonging of one or the other threat associated parameter to threat types, which is based on the GMM latent variables (the posterior probabilities that threat associated parameters belong to some GMM components), can be often made by means of contrast values $\theta_{n,k}$, which are $\theta_{n,k} \to 1$ or $\theta_{n,k} \to 0$ with a very high degree of compliance, e.g., for $\forall\theta_{n,4}$ in (10);
- self-adaptation of the proposed ML approach boils down to the fact that previous separation of subsets of threat associated parameters within $\mathbf{TAP}$ is not required and these parameters in the observed data can be not in the ascending order; subsets and typical a priori expected values of initial parameters are organized and adjusted automatically, which is also illustrated in Tables IV, V and VI for the modification of GMM parameters $\mathbf{M}$, $\mathbf{D}$, and $\mathbf{W}$.

TABLE III. Accuracy of Threat Types Detection

GMM component	Threat type	MAX_k	MIN_k	$MEAN_k$
k = 1	No threats	0.979	0.774	0.946
k = 2	Insignificant threat	0.738	0.559	0.662
k = 3	Moderate threat	1.000	0.684	0.961
k = 4	Significant threat	1.000	1.000	1.000

TABLE IV. Modification of GMM Parameters **M**

GMM component	Threat type	$M^{(0)}$	$M^{(80)}$
k = 1	No threats	1	13
k = 2	Insignificant threat	10	29
k = 3	Moderate threat	100	52
k = 4	Significant threat	1000	881

TABLE V. Modification of GMM Parameters **D**

GMM component	Threat type	$D^{(0)}$	$D^{(80)}$
k = 1	No threats	1000000	29.27
k = 2	Insignificant threat	1000000	12.32
k = 3	Moderate threat	1000000	306.08
k = 4	Significant threat	1000000	5849.19

TABLE VI. Modification of GMM Parameters **W**

GMM component	Threat type	$W^{(0)}$	$W^{(80)}$
k = 1	No threats	0.250	0.366
k = 2	Insignificant threat	0.250	0.099
k = 3	Moderate threat	0.250	0.368
k = 4	Significant threat	0.250	0.167

Among the peculiarities of the proposed ML approach, it should also be noted such peculiarities and relevant cases when indeterminate forms of the kind $0/0$ in the log-likelihood function $LLF(\mathbf{M}, \mathbf{D}, \mathbf{W}|\mathbf{TAP})$ can be obtained:

- some k-th empty subset is obtained when a relevant threat type is not detected, i.e., none of threat associated parameters TAP_n, $n = 1,2, \dots, N$, belong to some k-th GMM component ($W_k \rightarrow 0$ and $D_k \rightarrow 0$);
- some k-th subset contains only one threat associated parameter ($W_k \rightarrow 1/N$ and $D_k \rightarrow 0$);
- some k-th subset contains only the same threat associated parameters, e.g., TAP_{18} and TAP_{20} in the considered example ($W_k \rightarrow Y/N$, $Y \in \mathbb{N}$, $D_k \rightarrow 0$).

The above-mentioned peculiarities can be identified in practice as the "division by zero" cases. They require a special mathematical analysis of the of the log-likelihood function $LLF(\mathbf{M}, \mathbf{D}, \mathbf{W}|\mathbf{TAP})$ and its structure.

Conclusion

The paper deals with a problem of threat types detection in communication networks within a multivariate approach to cybersecurity problem, which take into account complex analysis of different threat types and their automatic unsupervised estimation and detection.

The GMM based ML approach for detection of threat types in communication networks is proposed an adjusted in the paper. A distinctive feature of the proposed ML approach is the use of the EM algorithm for detection of threat types.

The proposed ML approach uses as the observed data a vector of threat associated parameters, which are monitored parameters in communication network. These parameters can be associated with both features of the functioning of communication network and threat scenarios, e.g., delays and number of requests in cases of either specific routing or threats.

The detection of threat types in the proposed ML approach is performed by a forming of subsets of threat associated parameters using latent variables of GMM, which are by nature posterior probabilities that threat associated parameters are also associated with GMM components.

The example of simulation of the proposed ML approach, which deals with a detection of possible "no threats", "insignificant threat", "moderate threat", and "significant threat" cases in communication network, is shown and analyzed in the paper.

The main distinctive feature of the proposed ML approach is an automatic unsupervised organizing of subsets of threat associated parameters that are corresponding to threat types in communication networks. This organizing is also implemented without any previously setting of thresholds between ranges of these subsets, and using, in fact, only typical a priori expected values of threat associated parameters for each threat type. Another important feature of the proposed ML approach is the ability to estimate the accuracy of threat types detection. Other features and peculiarities of the proposed ML approach are also presented and analyzed in the paper.

Prospects for the development and practical use of the proposed ML approach are intelligent cybersecurity and data analysis systems for information and communication technologies and other fields, e.g., modern UAV control systems and technologies [32] etc.

References

[1] P. Thakur and G. Singh, "Cooperative spectrum monitoring in homogeneous and heterogeneous cognitive radio networks," in Spectrum Sharing in Cognitive Radio Networks: Towards Highly Connected Environments. Hoboken, NJ: Wiley Telecom, 2021, pp. 121-146. DOI: 10.1002/9781119665458.ch6

[2] M. K. Shehzad, M. W. Akhtar, and S. A. Hassan, "Performance of mmWave UAV-assisted 5G hybrid heterogeneous networks," in Autonomous Airborne Wireless Networks, M. A. Imran, Q. Abbasi, O. Onireti, and S. Ansari, Eds. Hoboken, NJ: Wiley-IEEE Press, 2021, pp. 97-118. DOI: 10.1002/9781119751717.ch6

[3] G. F. Elmasry, "IEEE standard for architectural building blocks enabling network-device distributed decision making for optimized radio resource usage in heterogeneous wireless access networks" in Dynamic Spectrum Access Decisions: Local, Distributed, Centralized, and Hybrid Designs. Hoboken, NJ: Wiley-IEEE Press, 2020, pp. 373-480. DOI: 10.1002/9781119573784.ch16

[4] O. Solomentsev, M. Zaliskyi, O. Shcherbyna, and O. Kozhokhina, "Sequential procedure of changepoint analysis during operational data

processing," in Proc. IEEE Microwave Theory and Techniques in Wireless Communications (MTTW), 2020, pp. 168-171. DOI: 10.1109/MTTW51045.2020.9245068

[5] Yu. Averyanova and E. Znakovskaja, "Weather hazards analysis for small UASs durability enhancement," in Proc. IEEE 6th International Conference on Actual Problems of Unmanned Aerial Vehicles Development (APUAVD), 2021, pp. 41-44. DOI: 10.1109/APUAVD53804.2021.9615440

[6] V. J. Larin and E. E. Fedorov, "Combination of PNN network and DTW method for identification of reserved words, used in aviation during radio negotiation", Radioelectronics and Communications Systems, vol. 57, no. 8, pp. 362-368, 2014. DOI: 10.3103/S0735272714080044

[7] I. Ostroumov and N. Kuzmenko, "Configuration analysis of European navigational aids network," in Proc. Integrated Communications Navigation and Surveillance Conference (ICNS), 2021, pp. 1-9. DOI: 10.1109/ICNS52807.2021.9441576

[8] I. Ostroumov, K. Marais, and N. Kuzmenko, "Aircraft positioning using multiple distance measurements and spline prediction," Aviation, vol. 26, no. 1, pp. 1-10, 2022. DOI: 10.3846/aviation.2022.16589

[9] Yu. Averyanova, A. Rudiakova, and F. Yanovsky, "Aircraft trajectories correction using sharing operative meteorological radar information," in Proc. 21st International Radar Symposium (IRS), 2020, pp. 256-259. DOI: 10.23919/IRS48640.2020.9253799

[10] I. Ostroumov and N. Kuzmenko, "Statistical analysis and flight route extraction from automatic dependent surveillance-broadcast data," in Proc. Integrated Communications Navigation and Surveillance Conference (ICNS), 2022, pp. 1-9. DOI: 10.1109/ICNS54818.2022.9771515

[11] O. A. Sushchenko, Y. M. Bezkorovainyi, and V. O. Golitsyn, "Modelling of microelectromechanical inertial sensors," in Proc. IEEE 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), 2019, pp. 23-27. DOI: 10.1109/CADSM.2019.8779286

[12] S. Frattasi and F. Della Rosa, "Application areas of positioning," in Mobile Positioning and Tracking: From Conventional to Cooperative Techniques, 2nd ed. Hoboken, NJ: Wiley-IEEE Press, 2017, pp. 11-42. DOI: 10.1002/9781119068846.ch2

[13] D. M. Thomas, N. Pandey, V. K. Shukla, and A. V. Singh, "Attack vectors and susceptibilities of the modbus in TCP/IP model," in Proc. 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1-5. DOI: 10.1109/ICRITO51393.2021.9596460

[14] Y. Qian, F. Ye, and H. Chen, "Security in 5G wireless networks," in Security in Wireless Communication Networks. Hoboken, NJ: Wiley-IEEE Press, 2021, pp. 281-310. DOI: 10.1002/9781119244400.ch14

[15] S. Lange, S. Schwarzmann, M. Gaji´c, T. Zinner, and F. A. Kraemer, "AI in 5G networks: challenges and use cases," in Communication Networks and Service Management in the Era of Artificial Intelligence and Machine Learning, N. Zincir-Heywood, M. Mellia, Y. Diao, Eds. Hoboken, NJ: Wiley-IEEE Press, 2021, pp. 101-122. DOI: 10.1002/9781119675525.ch5

[16] O. A. Sushchenko, Y. M. Bezkorovayniy, and V. O. Golitsyn, "Processing of redundant information in airborne electronic systems by means of neural networks," in Proc. IEEE 39th International Conference on Electronics and Nanotechnology (ELNANO), 2019, pp. 652-655. DOI: 10.1109/ELNANO.2019.8783394

[17] O. A. Sushchenko, Y. M. Bezkorovainyi, and V. O. Golitsyn, "Fault-tolerant inertial measuring instrument with neural network," in Proc. IEEE 40th International Conference on Electronics and Nanotechnology (ELNANO), 2020, pp. 797-801. DOI: 10.1109/ELNANO50318.2020.9088779

[18] A. Halbouni, T. S. Gunawan, M. H. Habaebi, M. Halbouni, M. Kartiwi, and R. Ahmad, "Machine learning and deep learning approaches for cybersecurity: a review," IEEE Access, vol. 10, pp. 19572-19585, 2022. DOI: 10.1109/ACCESS.2022.3151248

[19] S. Hariharan, A. Velicheti, A. S. Anagha, C. Thomas, and N. Balakrishnan, "Explainable artificial intelligence in cybersecurity: a brief review," in Proc. 4th International Conference on Security and Privacy (ISEA-ISAP), 2021, pp. 1-12. DOI: 10.1109/ISEA-ISAP54304.2021.9689765

[20] F. Chamroukhi and B. T. Huynh, "Regularized maximum-likelihood estimation of mixture-of-experts for regression and clustering," in Proc. International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1-8. DOI: 10.1109/IJCNN.2018.8489670

[21] X. Li, S. Liu, M. Liu, and Z. Tian, "Maximum likelihood based pairwise clustering," in Proc. 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011, pp. 1147-1151. DOI: 10.1109/FSKD.2011.6019653

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal statistical society. Series B, vol. 39, no. 1, pp. 1-38, 1977.

[23] D. Yu and L. Deng, "Gaussian mixture models," in Automatic Speech Recognition. A Deep Learning Approach. London: Springer-Verlag, 2014, pp. 13-21. DOI: 10.1007/978-1-4471-5779-3_2

[24] T. Huang, H. Peng, and K. Zhang, "Model selection for Gaussian mixture models," Statistica Sinica, vol. 27, pp. 147-169, 2017. DOI: 10.5705/ss.2014.105

[25] M. R. Gupta and Y. Chen, "Theory and use of the EM algorithm," Foundations and Trends® in Signal Processing, vol. 4, no. 3, pp. 223-296, 2011. DOI: 10.1561/2000000034

[26] N. Vlassis and A. Likas, "A greedy EM algorithm for Gaussian mixture learning," Neural Processing Letters, vol. 15, pp. 77-87, 2002. DOI: 10.1023/A:1013844811137

[27] K. Bekiroglu, A. Tekeoglu, B. Andriamanalimanana, S. Sengupta, C. Chiang, and Jorge Novillo, "Hankel-based unsupervised anomaly detection," in Proc. American Control Conference (ACC), 2020, pp. 5139-5144. DOI: 10.23919/ACC45564.2020.9147583

[28] M. Zaliskyi, Yu. Petrova, M. Asanov, and E. Bekirov, "Statistical data processing during wind generators operation," International Journal of Electrical and Electronic Engineering & Telecommunications, vol. 8, no. 1, pp. 33-38, 2019. DOI: 10.18178/ijeetc.8.1.33-38

[29] O. Solomentsev and M. Zaliskyi, "Method of sequential estimation of statistical distribution parameters in control systems design," in Proc. IEEE 3rd International Conference on Methods and Systems of Navigation and Motion Control (MSNMC), 2014, pp. 135-138. DOI: 10.1109/MSNMC.2014.6979752

[30] V. Larin, N. Chichikalo, K. Larina, and A. Shcherban, "Algorithm for processing of informative and influencing factors in UAV Battery discharge management system," in Proc. IEEE 6th International Conference on Actual Problems of Unmanned Aerial Vehicles Development (APUAVD), 2021, pp. 130-134. DOI: 10.1109/APUAVD53804.2021.9615406

[31] Yu. Averyanova, A. Rudiakova, and F. Yanovsky, "Drop deformation estimate with multi-polarization radar," International Journal of Microwave and Wireless Technologies, vol. 12, no. 9, pp. 870-877, 2020. DOI: 10.1017/S1759078720000732

[32] V. J. Larin, "Development of measurement-based feedback on 3-D coil non-invasive transducer in UAV's rotation gear control unit," in Proc. IEEE International Conference Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD), 2015, pp. 270-273. DOI: 10.1109/APUAVD.2015.7346617

Unsupervised Detection of Anomalous Running Patterns Using Cluster Analysis

Ivan Ursul
Faculty of Applied Mathematics and Informatics
Ivan Franko National University of Lviv
Lviv, Ukraine
ivan.ursul@lnu.edu.ua

Andriy Pereymybida
Faculty of Applied Mathematics and Informatics
Ivan Franko National University of Lviv
Lviv, Ukraine
andrii.pereimybida@lnu.edu.ua

Abstract — **Anomaly detection is an important problem in various domains, such as user analysis, network intrusion detection, fraud detection and system monitoring. This paper provides a comprehensive review of anomaly detection algorithms for clustering applications. Hence, the paper provides a detailed analysis of various clustering techniques, including distance based, hierarchical, and density based. They also discussed using ensemble techniques and outlier detection methods to improve anomaly detection accuracy on this data set. The authors have applied them to a customer data set to detect anomalies. The authors discussed different cluster-based algorithms and compared their performance based on scalability, precision, recall, and f1 score metrics. Moreover, this work highlights the challenges of cluster-based anomaly detection, such as selecting the appropriate number of clusters, dealing with high-dimensional data, and handling imbalanced datasets. Finally, the authors provided insights into addressing these challenges and discussed future research directions.**

Keywords — clusters, anomaly detection, unsupervised learning

I. Introduction

In today's world, data grows exponentially; scientists predict that the 'tsunami of data is coming in recent years '[1], as more users will join the global internet, and more signals will be gathered about them on a daily basis. This torrent of information will not only come from an increase in the number of users [2] but also from the expanding capabilities of our technological infrastructure [3], including Internet of Things (IoT) devices [2], cloud-based platforms [4], and digital service providers. The data is growing not just in volume but also in complexity and diversity.

As our world becomes increasingly digital and interconnected, the speed and volume of data will far exceed human capabilities for processing and analysis [5]. With more data in the coming years, we expect to see an increased demand for automated data analysis [6]. As a result, machine learning and artificial intelligence are anticipated to play a more prominent role in extracting insights and making predictions from large datasets.

Existing methods of data analysis are semi-automated; they require a certain amount of human interaction for doing the pre-processing: filtering the expected data, and finding the data that does not fit into any of the known patterns [7]. However, as we move into the future, there is a growing need to fully automate these steps to manage the increased scale and complexity of the data [8]. In this regard, research is underway to develop more sophisticated algorithms and models that can handle these tasks more efficiently and accurately. One of the significant milestones of fully automated data analysis is automated outlier detection. The problem with finding outliers or anomalies in data is that 'we don't know what we don't know '[9]. Moreover, automated systems are prone to misuse, as they can be built to serve one set of functions, while later being used by unauthorized users for completely different purposes. The problem of limited knowledge of the data perfectly describes the problem of unsupervised learning; finding an outlier or anomaly if we do not know how it looks. However, the advancements in unsupervised learning and artificial intelligence hold promise for improving anomaly detection and managing outlier or anomalous data [10].

It is worth mentioning that not all anomalies are bad [11]: for example, anomaly detection can be used to identify customers who exhibit unique behaviors or preferences that may not be captured by traditional segmentation methods [12]. This is an important area of research, as identifying these unique customer behaviors can provide businesses with valuable insights into customer preferences and behaviors that can be leveraged for business growth.

Cluster-based anomaly detection [13] could be an answer to the problems mentioned. The idea is to cluster similar data points and identify outliers that do not fit perfectly into the specific cluster. Depending on the type of clustering algorithm, a different method for detecting an outlier is used. This science paper is focused on partition-based, density-based and hierarchical clustering algorithms. The main challenge in cluster-based anomaly detection is finding an approach that could work well for large-scale data, will be insensitive to the choice of input parameters, and can provide insightful information about anomalies. To deal with this challenge, we are exploring a number of different strategies, including but not limited to, the use of ensemble methods, feature engineering, and optimized parameter tuning. Our research also considers the trade-off between computation time and accuracy, which is a critical factor when dealing with large-scale data.

In the following sections, this paper will delve deeper into the topic of anomaly detection using machine learning techniques. Section II will provide a review of various clustering algorithms used for anomaly detection. It will detail the advantages, disadvantages, and typical use cases for each

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

type of algorithm. Section III will describe the dataset used in the study and the experimental setting. This section will also elaborate on the data preprocessing steps and any challenges faced during this phase. Section IV will present the results and analysis of the evaluation of the machine learning algorithms. It will provide a comprehensive discussion of the performance of each algorithm and the reasons for their respective performances. Finally, Section V will conclude the findings and their implications, offering future directions for research and potential applications of the findings.

II. Review Of Clustering Algorithms For Anomaly Detection

This section briefly reviews three major clustering algorithms commonly used for anomaly detection: Partition-based clustering, Density-based clustering, and Hierarchical-based clustering. This review will focus on the intuition, strengths, and limitations of each algorithm.

A. Partition-based clustering

Partition-based clustering is a commonly used clustering algorithm for anomaly detection. Partition-based algorithms such as K-Means [15] involve dividing the dataset into K clusters based on the similarity of data points. While these algorithms are effective in identifying patterns and grouping similar data points, they may not always be able to detect anomalies. One approach to identify anomalies in K-Means [15] clustering is to define a threshold value based on the within-cluster sum of squares (WCSS) for each cluster. The WCSS measures the sum of the squared distances between each data point and the centroid of its assigned cluster. To detect anomalies, we can calculate the WCSS for each cluster and define a threshold value above the average WCSS for all clusters. Any data point with a WCSS value above this threshold can be considered an anomaly. Formally, let $C_1,C_2,...,C_k$ be the K clusters generated by the K-Means [15] algorithm, and let $S_1,S_2,...,S_k$ be the corresponding sum of squares for each cluster. The average WCSS, denoted by S_{avg}, can be calculated as:

$$S_{avg} = \frac{(S_1,S_2,...,S_k)}{K} \tag{1}$$

Next, we define a threshold value, T, a multiple of the standard deviation of the sum of squares for each cluster. The threshold value can be calculated as:

$$T = S_{avg} + k * SD[S_1, S_2, ..., S_k], \tag{2}$$

where SD is the standard deviation function, and k is a constant that determines the sensitivity of anomaly detection.

Finally, any data point whose WCSS value is above the threshold value T can be identified as an anomaly. This approach can identify anomalies in partition-based algorithms such as KMeans [15], allowing for more robust and comprehensive data analysis. The Visualization of K-Means clustering is provided in figure 1.

B. Density-based clustering

Density-based clustering is another commonly used clustering algorithm for anomaly detection. To detect anomalies using DBSCAN [16], two parameters are used: epsilon (ϵ) and minimum points (MinPts). Epsilon is the maximum distance between two data points for them to be considered neighbors, and MinPts is the minimum number of neighboring data points for a data point to be considered a core point. A core point is a data point with at least MinPts neighboring data points within a distance of ϵ. A border point is a data point that does not have enough neighboring data points to be a core point but is within a distance of ϵ from a core point. A noise point is a data point not a core or a border point.

DBSCAN [16] starts by selecting a random core point and finding all neighboring core points within a distance of ϵ. These core points are then merged into a cluster. The process is repeated until all core points have been assigned to clusters. Border points are assigned to the cluster of their nearest core point, and noise points are not assigned to any cluster. After clustering the data points, anomalies can be identified as noise points or data points that belong to clusters with a small number of data points.

The problem of density-based clustering like DBSCAN [16] for anomaly detection can be formulated as a binary classification problem, where the objective is to classify each data point as either an anomaly or a normal data point. DBSCAN [16] can be used in various applications, such as fraud detection, intrusion detection, and fault diagnosis.

Let $C = C_1, C_2, ..., C_k$ be the set of all clusters identified by DBSCAN [16], where k is the total number of clusters. Let $S = \{S_1,S_2,...,S_m\}$ be the set of all core points in the dataset, where m is the total number of core points. Let $B = \{B_1,B_2,...,B_l\}$ be the set of all border points in the dataset, where l is the total number of border points. Then, the objective function of DBSCAN [16] can be written as:

$$C = \{C_1,C_2,...,C_k\} = \{S_1 \cup B_1, S_2 \cup B_2,...,S_k \cup B_k\} \tag{3}$$

Each C_i is a cluster, defined as the union of a core point and its corresponding border points. The number of clusters k is unknown in advance and may vary depending on the data and the chosen values of ϵ and MinPts. DBSCAN [16] is a powerful tool for anomaly detection that can be used to identify clusters of data points with a high density of neighboring data points and detect anomalies with a low density of neighboring data points.

C. Hierarchical-based clustering

Hierarchical clustering is a commonly used unsupervised machine learning algorithm for anomaly detection. The problem of hierarchical clustering for anomaly detection can be formally defined above in II-B To detect anomalies using hierarchical clustering, a distance metric is used to measure the similarity or dissimilarity between data points. One common distance metric used for anomaly detection is the Mahalanobis distance, as defined in the problem of agglomerative hierarchical clustering.

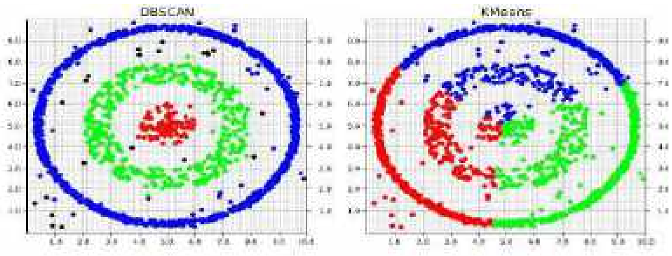


Fig. 1. The Visualization of Comparison of K-Means Clustering and DBSCAN image sources: [21]

Hierarchical clustering can be performed using two approaches: agglomerative clustering and divisive clustering. Agglomerative clustering starts with each data point as a separate cluster and iteratively merges the closest clusters until a stopping criterion is met. Divisive clustering starts with all data points in a single cluster and recursively splits it into smaller clusters until a stopping criterion is met.

The similarity or dissimilarity between two clusters is measured using a linkage criterion, which defines the distance between two clusters. Different linkage criteria can be used for anomaly detection, such as single linkage, complete linkage, average linkage, and Ward's linkage. The linkage criterion used can affect the structure of the resulting hierarchy and the quality of the clustering.

After clustering the data points, anomalies can be identified as data points that do not belong to any of the clusters or belong to clusters with a small number of data points. The threshold for determining the size of a cluster can be set based on domain knowledge or using statistical methods such as the Elbow method or the Silhouette method. The objective function of hierarchical clustering for anomaly detection can be written as:

$$min \sum_{C_i \in C} \sum_{x_i, x_j \in c_i} d(x_i, x_j) \quad (4)$$

$C_i \in C$ $x_i,x_j \in C_i$ where k is the number of clusters, C_i is the i-th cluster, and $d(x_i, x_j)$ is the distance between data points x_i and x_j. The goal is to minimize the within-cluster dissimilarity and maximize the between-cluster dissimilarity.

III. Methodology

In this section, the methodology of the proposed study will be explained in detail, including the dataset used, the experimental setting, and the error minimization techniques employed. This information will provide a comprehensive overview of the study's approach and help contextualize the results obtained from the evaluation of machine learning algorithms for anomaly detection.

A. Datasets used.

In this study, we used a dataset of running activity data recorded by a wearable device. The dataset includes the following columns: datetime, athlete, distance, duration, gender, age group, country, and major. The datetime column indicates the date and time of the running activity, while the athlete column identifies the individual who performed the activity. The distance and duration columns indicate the distance and duration of the running activity, respectively. The gender and age group columns provide demographic information about the athlete, while the country and major columns provide additional background information. The dataset is balanced, with a total of 10,703,690 running activities recorded. It also contains a subset of false activity labels, where individuals recorded an activity without actually making the running session.

The pace(duration/distance) distribution chart helps us understand that the median pace is around 5:20, which is slightly higher than the average pace around the recreational runners, according to multiple studies. The analysis of the distance shows that the majority of running activities are within 5-10 kilometers range. Additionally, the dataset includes data about top performers, who may exhibit different running patterns than the general population. It is worth noting that the duration and distance of the running activities are quite imbalanced, with a wide range of values recorded. Specifically, the dataset includes running sessions for marathons, and half-marathons, 10 and 5 kilometres were tracked. The balanced nature of the dataset, the variation of distances, along with the presence of false activity labels and top performer data, pose significant challenges for anomaly detection.

B. Definition of anomaly

In the context of this study, an anomaly refers to a running activity that deviates significantly from the expected or normal pattern of behavior. The expected pattern of behavior is determined based on the characteristics of the dataset, including the distribution of running distances, durations, and paces.

C. Experiment Setting

We conducted a data analysis to determine the expected pattern of behavior for our dataset. Based on this analysis, we define an anomalous running pattern as one of the following:

- Top performers with the smallest pace (99p): In our dataset, some athletes may be top performers who exhibit a smaller pace than the average athlete, thus their running distances may be different from the distances of average performers. These athletes may exhibit different running patterns, which can be identified as anomalies.
- Low performers with an unusually large pace: Similarly, some athletes in the dataset may be low performers who exhibit an unusually large pace. These athletes may also exhibit different running patterns, which can be detected as anomalies.
- Different activity: In addition to false activity labels, the dataset may also include other activities, such as hiking or weight lifting, which can be distinguished from running activities. These activities may be detected as anomalies in the dataset.

D. Error minimization techniques

In this exploration, we concentrated on conducting anomaly detection in cluster-based operations by employing robust ways to minimize errors and enhance the overall performance of our clustering algorithms. Two essential ways were applied to achieve these are feature selection and parameter tuning.

Feature selection aimed to identify a subset of applicable features, $F' \subseteq F$, where F denotes the original set of features, to capture the beginning structure of the data and effectively separate between normal and anomalous cases. We employed various ways, like correlation analysis and collective information, to quantify the significance of each feature in the environment of anomaly discovery.

Also, we also performed expansive parameter tuning to optimize the hyperparameters for each clustering algorithm under disquisition. Let θ denote the set of hyperparameters for a given algorithm and let $L(\theta)$ represent the loss function that quantifies the divagation between the algorithm's prognostications and the true markers. Our ideal was to find the optimal hyperparameters θ^{*} that minimize the loss function:

$$\theta^{*} = argmin\theta L(\theta) \tag{5}$$

IV. Results and Analysis

The main idea of the research was to compare different types of clustering algorithms. Precision, recall and F1 score metrics were used to compare the efficiency of clustering algorithms. Partition-based, Density-based and hierarchical clustering algorithms were picked for comparison.

TABLE I. PERFORMANCE COMPARISON OF CLUSTERING ALGORITHMS

Algorithm	Precision	Recall	F1 Score
K-Means [15]	0.98	0.053	0.1
DBSCAN [16]	0.92	0.99	0.95
HDBSCAN [17]	0.88	0.05	0.1
Optics [18]	0.95	0.66	0.78
Local Outlier Factor [19]	0.84	0.09	0.16
Mean Shift [20]	0.96	0.05	0.09

In examining the performance of various clustering algorithms, as presented in Table I, discernible differences in precision, recall, and F1 scores were identified. The K-Means algorithm [15], while it exhibited a high precision of 0.98, yielded rather low recall and F1 scores, at 0.053 and 0.1 respectively. This suggests that while K-Means was proficient in accurately classifying relevant instances, it struggled with retrieving all relevant instances.

DBSCAN [16], on the other hand, outperformed K-Means across all measures. It maintained a high precision of 0.92, and considerably better recall and F1 scores, at 0.99 and 0.95 respectively. This illustrates that DBSCAN was effective not only in accurately classifying instances, but also in retrieving most of the relevant instances. While the HDBSCAN algorithm [17], demonstrated an average precision of 0.88, but much like K-Means, struggled with recall, exhibiting a low score of 0.05. Consequently, its F1 score was also quite low, at 0.1.

On the other hand, the Optics algorithm [18] showed a solid performance, with a precision of 0.95 and a recall of 0.66. These results culminated in a high F1 score of 0.78. The Local Outlier Factor [19] algorithm, however, displayed lower precision, recall, and F1 scores of 0.84, 0.09, and 0.16 respectively. The Mean Shift algorithm [20] displayed high precision, like K-Means, at 0.96. Yet, it suffered from a low recall of 0.05 and consequently a low F1 score of 0.09.

To summarize, the DBSCAN algorithm, in this evaluation, outperformed all others across precision, recall, and F1 score.

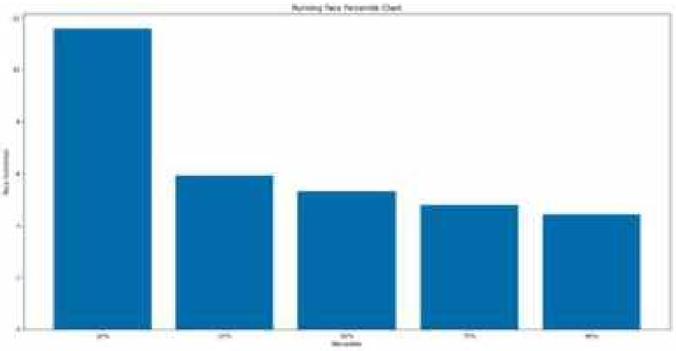


Fig. 2. Running Pace Percentile Chart

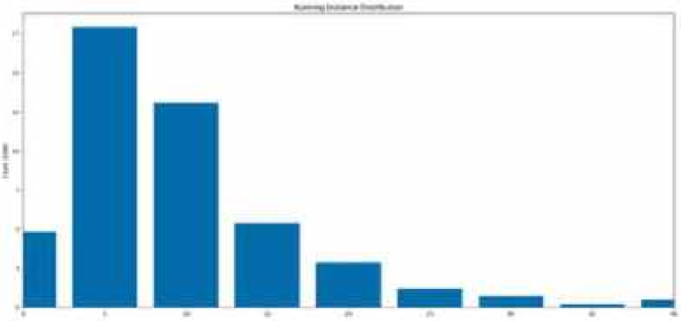


Fig. 3. Running Distance Distribution

Nevertheless, the choice of algorithm will inherently depend on the specific needs and constraints of the application at hand. the pictorial representation of the result can also be seen in the figure 2 and 3.

Through an iterative process involving methods similar to grid hunting and Bayesian optimization, we linked the stylish set of hyperparameters that yielded optimal clustering results for each algorithm. By integrating these two methods—feature selection and parameter tuning—we achieved more accurate and dependable anomaly discovery performance across clustering algorithms. The methodical approach of incorporating these strategies allowed us to reduce errors and enhance the effectiveness of our chosen algorithms in the environment of cluster-based anomaly detection algorithms.

V. Conclusion

In this study, we proposed a cluster-based unsupervised anomaly detection method for identifying anomalous running patterns in a running activity dataset. Our method was able to detect anomalous running patterns, including those exhibited by top performers, low performers, and other types of physical activities. We evaluated the performance of several clustering and anomaly detection algorithms, including DBSCAN [16], HDBSCAN [17], OPTICS [18], Mean Shift [20], Local Outlier Factor [19], and K-Means [15]. Our results showed that DBSCAN [16] exhibited the best performance in terms of the quality of anomaly detection. DBSCAN [16] was also computationally efficient for large datasets like the one used in our research.

We found that HDBSCAN [17], OPTICS [18], Mean Shift [20], Local Outlier Factor [19], and K-Means [15] also showed promising results in detecting anomalous running patterns. However, we noted that the Local Outlier Factor [19] was computationally slow on our large dataset. Our study provides valuable insights into the detection of anomalous running patterns using cluster analysis. Our method can be used to identify not only the expected patterns of behavior but also the unexpected ones like incorrectly tracked activities or other types of activities. These insights can be used to improve individual and group running performance and health outcomes.

However, it is important to note that our method cannot detect cheating activity due to the low dimensionality of the dataset and the limited feature selection. One option to detect cheating could be using telemetry data from wearable devices in combination with time-series anomaly detection techniques. Overall, our study demonstrates the potential of unsupervised anomaly detection methods for identifying anomalous running patterns in large and complex datasets. We recommend further investigation into the use of these methods for analyzing physical activity data in other contexts and the development of more advanced techniques for detecting cheating in running activity data.

VI. Future Work

In future research, we plan to provide a comprehensive systematic review of advancements in unsupervised anomaly detection methods for various application domains, including sports, healthcare, and finance. This review will aim to synthesize the latest findings, identify challenges, and propose new directions in the development and application of unsupervised anomaly detection techniques. We believe that our study and the proposed future work will help to advance the state-of-the-art in unsupervised anomaly detection and contribute to its practical applications. Additionally, we will investigate the potential of incorporating more advanced features and data sources, such as telemetry data from wearable devices and other sensor-based information, to improve the accuracy and robustness of our anomaly detection method. This may enable us to detect more subtle anomalies, such as cheating activity or early signs of injury, which could be overlooked by our current approach due to the limited dimensionality of the dataset. Furthermore, we plan to explore the application of deep learning techniques, such as autoencoders and variational autoencoders, for unsupervised anomaly detection. These methods have shown promise in learning complex, high-dimensional feature representations and detecting subtle patterns in data, which may improve our ability to identify anomalous running patterns and other anomalies in physical activity data.

References

[1] Andrae, A.S., 2019. Prediction studies of electricity use of global computing in 2030. International Journal of Science and Engineering Investigations, 8(86), pp.27-33.

[2] Singh, D. (2023). Internet of Things. Factories of the Future: Technological Advancements in the Manufacturing Industry, 195227.

[3] Broo, D. G., & Schooling, J. (2023). Digital twins in infrastructure: definitions, current practices, challenges and strategies. International Journal of Construction Management, 23(7), 12541263.

[4] Mao, Z., Yuan, Q., Li, H., Zhang, Y., Huang, Y., Yang, C., ... & Ma, H. (2023). CAVE: a cloud-based platform for analysis and visualization of metabolic pathways. Nucleic Acids Research, gkad360.

[5] Miller, A. (2019). The intrinsically linked future for human and Artificial Intelligence interaction. Journal of Big Data, 6(1), 38.

[6] Laato, S., Mantym¨ aki, M., Islam, A. N., Hyrynsalmi, S., &¨ Birkstedt, T. (2023). Trends and Trajectories in the Software Industry: implications for the future of work. Information Systems Frontiers, 25(2), 929-944.

[7] Sushentsev, N., Moreira Da Silva, N., Yeung, M., Barrett, T., Sala, E., Roberts, M., & Rundo, L. (2022). Comparative performance of fully-automated and semi-automated artificial intelligence methods for the detection of clinically significant prostate cancer on MRI: a systematic review. Insights into imaging, 13(1), 1-17.

[8] Armenatzoglou, N., Basu, S., Bhanoori, N., Cai, M., Chainani, N., Chinta, K., ... & Terry, D. (2022, June). Amazon Redshift re-invented. In Proceedings of the 2022 International Conference on Management of Data (pp. 2205-2217).

[9] Schwartz, T. (2011). We don't know what we don't know. Book review, Harvard Business Review.

[10] Lindemann, B., Maschler, B., Sahlab, N., & Weyrich, M. (2021). A survey on anomaly detection for technical systems using LSTM networks. Computers in Industry, 131, 103498.

[11] Saleem, T. J., & Chishti, M. A. (2021). Deep learning for the internet of things: Potential benefits and use-cases. Digital Communications and Networks, 7(4), 526-542.

[12] Chandola, V., Banerjee, A., & Kumar, V. (2010). Anomaly detection for discrete sequences: A survey. IEEE transactions on knowledge and data engineering, 24(5), 823-839.

[13] Bigdeli, E., Mohammadi, M., Raahemi, B., & Matwin, S. (2017). A fast and noise resilient cluster-based anomaly detection. Pattern Analysis and Applications, 20, 183-199.

[14] Wang, D., Liao, Q.V., Zhang, Y., Khurana, U., Samulowitz, H., Park, S., Muller, M. and Amini, L., 2021. How much automation does a data scientist want?. arXiv preprint arXiv:2101.03970.

[15] Hartigan, J.A. and Wong, M.A., 1979. Algorithm AS 136: A kmeans clustering algorithm. Journal of the royal statistical society. series c (applied statistics), 28(1), pp.100-108.

[16] Hinneburg, A., 1996. A density based algorithm for discovering clusters in large spatial databases with noise. In KDD Conference, 1996.

[17] Campello, R.J., Moulavi, D. and Sander, J., 2013. Density-based clustering based on hierarchical density estimates. In Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17,

[18] 2013, Proceedings, Part II 17 (pp. 160-172). Springer Berlin Heidelberg.

[19] Ankerst, M., Breunig, M.M., Kriegel, H.P. and Sander, J., 1999. OPTICS: Ordering points to identify the clustering structure. ACM Sigmod record, 28(2), pp.49-60.

[20] Bhatt, V., Dhakar, M. and Chaurasia, B.K., 2016. Filtered clustering based on local outlier factor in data mining. International Journal of Database Theory and Application, 9(5), pp.275-282.

[21] Cheng, Y., 1995. Mean shift, mode seeking, and clustering. IEEE transactions on pattern analysis and machine intelligence, 17(8), pp.790-799.

[22] Page, J.T., Liechty, Z.S., Huynh, M.D. and Udall, J.A., 2014. BamBam: genome sequence analysis tools for biologists. BMC Research Notes, 7(1), pp.1-5.

Nvidia Jetson NanoPlatform Using for Accelerating Image Recognition

Dmytro Myronyuk
Dept. of Radiophysics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
myronyukdmytro@gmail.com

Bohdan Blagitko
Dept. of Radiophysics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
blagitko@gmail.com

***Abstract* — This article analyzes modern methods of image recognition using CUDA and TensorRT hardware and software technologies. Method accelerating the execution of ready-made mathematical models of neural networks to image recognition implemented on the Nvidia Jetson Nano Platform. Proprietary mathematical neural network fast image recognition on the YOLOv5 model was created. It's transformed and optimized by CUDA and TensorRT tools. A performance-implemented neural network optimized and non-optimized versions comparison make.**

Keywords* — *mathematical modeling, computer vision, image recognition

I. INTRODUCTION

One of the main areas of development of computer vision systems is systems based on specialized low-power mobile platforms. Using such platforms makes it possible to expand software systems using artificial intelligence elements and to improve the general characteristics of finished software and hardware systems. The current level of development of computer vision systems makes it possible to use them on platforms with less computing power without significant loss in accuracy for high-quality performance of final tasks. From the point of view of the software component, in order to run mathematical neural network models of computer vision on such platforms, it is necessary that the models have a high level of optimization to run with sufficient speed and accuracy.

The mobile platform models, as a rule, have a single-stage architecture with maximally simplified branches of operators in the network graph. An example of such detection systems can be considered the MobileNet and YOLO architectures. It has quietly different operating principles but can be used equally for pattern recognition on low-power platforms. Fig. 1 shows the general concept of the YOLOv4 architecture [8].

The YOLO image recognition algorithm is a fast and fair accurate solution for working on platforms of different types, architectures, and power. The algorithm core version has been supported by Ultralytics. It offers initial pre-trained models for research and commercial use. The approach uses zoning of the images in one pass (by dividing them into squares and working with each one separately). The object per zone determination is carried out by estimating the distance from the center of each square to the object zone (the anchor superimposed on the entire image). It is how the selection and assessment of the best areas for finding objects is carried out. As of the time of research, the architecture has gone through 8 generations of improvements and is one of the main ones for object detection on mobile platforms.

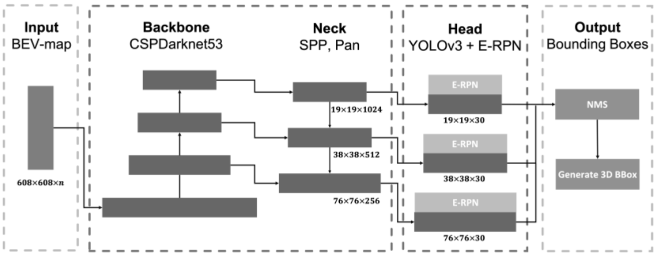


Fig. 1. General concept of YOLOv4 architecture [10].

A model based on yolov5s was used for the study. This includes the following subsets of layers:

- Backbone: CSP-DarkNet53;
- Neck: SPFF;
- Head: YOLOv3.

The study of the speed of the models was carried out in two different modes: on the standard pre-trained yolov5s models and on the user model for use in the recognition algorithm of the mechanical manipulator model.

II. TENSORRT SOFTWARE FRAMEWORK IS ONE OF THE MAIN TOOLS FOR THE ACCELERATION OF MACHINE LEARNING MODELS ON THE MOBILE PLATFORM NVIDIA

TensorRT software framework is one of the main means for accelerating the process of executing ready-made models on devices manufactured by Nvidia. The tool is able to work with popular software frameworks for training models — TensorFlow, PyTorch, and MXNet. The main purpose is the ready-made trained model's acceleration and optimization on Nvidia platforms.

TensorRT has a usage model that consists of two stages:

- The first phase is model optimization for a certain platform. As a rule, it is performing locally. It consists of providing a model to specialized parsers for certain software frameworks. It performs analysis and actual optimization according to the recommended parameters of a specific platform where the current model is running. The easiest way is to convert the trained model into the ONNX format. It supports by the vast majority of modern machine-learning software frameworks. Then builds the model based on TensorRT optimization primitives. The next subsection of the first stage is the setting of input and output tensors for the model. Next, the model is building for execution on TensorRT primitives. In this stage, the user can control the maximum reduction in accuracy. It should be the result of the optimization and the balance between the program execution speed and its accuracy. The model builder for TendorRT supports the post-quantization of models with

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

post-calibration to reduce memory usage and increase execution speed. The floating point quantization to half-precision (float16) is supported currently, as well as integer quantization int8, according to the framework documentation [2].

- The second phase is execution on the specialized TensorRT engine, which supports splitting large GPUs into different threads and execution on all platforms. The additional processing tools (such as DeepStream, a tool for scripting and constructing models) allow the application to run at near real-time speeds. The manufacturer declares an acceleration of up to 32 times [2] compared to conventional motors. The main technology used by the engine to execute the program on graphics accelerators is the CUDA software-hardware framework, which was used in previous studies [3]. Nvidia CUDA is a software-hardware platform for computing on GPUs. It was developed and maintained by Nvidia. It contains a great set of tools for performing a variety of general computing tasks, as well as packages for parallelization and solving specific tasks. It presents in the C programming language extension as a software form. The proprietary nvcc compiler uses to translate the code from this extension. It was created on the basis of the open Open64 compiler.

The key features of CUDA:

- The main unified solution for performing a common task using Nvidia graphics processors.
- A large set of supported solutions.
- A large set of standard libraries for numerical analysis (including BLAS and FFT).
- Optimized for efficient data exchange between CPU and GPU.
- Interaction with graphic API OpenGL and DirectX.
- Possibility of low-level development.
- Support for a wide range of operating systems.
- High documentation and a great set of code examples for beginners.

III. NVIDIA SOFTWARE PRODUCTS USED TO TRAIN THE MODELS

A. Training data and process parameters used

To train the neural network model as the main algorithm for recognizing the operation of the manipulator, we used our own data set containing 300 images of 4 categories, which is training for "object" recognition. TABLE I shows the general characteristics of the data set.

TABLE I. CHARACTERISTICS OF THE TRAINING DATA SET

#	***Characteristic***	***Value***
1	Number of unique images	300 + 30 (training and validation dataset)
2	Number of classes	4
3	Augmentation	Rotations, affine transformations, scaling
4	Size	608x608x3RGB

The choice of image size is justified by the small size of the input image, which increases the model speed on the mobile hardware platform without significant loss in recognition accuracy. TABLE II shows the training parameters.

TABLE II. EDUCATIONAL PARAMETERS

#	***Parameter***	***Value***
1	Backbone	CSP-DarkNet53
2	Learning strategy	Transfer Learning
3	Strategy to stop learning	Step limit

B. Use of stationary hardware resources

For training, standard pre-trained models use on the ImageNet dataset from the standard models package provided by the Ultralytics team together with the raw code of the YOLO approach [5] for the Pytorch software framework. The validation data set photos are used for testing also.

A laptop based on the seventh-generation Intel Core i5 processor was used as a hardware platform. TABLE III shows the characteristics of the platform.

TABLE III. THE HARDWARE PLATFORM CHARACTERISTICS FOR TRAINING

#	***Parameter***	***Value***
1	Processor	Intel Core i5 7300 HQ 2.5-3.4 GHz
2	RAM	8 GB
3	Storage device	SSHD TOSHIBAMQ02ABD1 1 Tb
4	Graphics processor	Nvidia GTX 1050 4 GB
5	CUDA	v. 11.8.89
6	Driver version	v.520.56.06
7	Operating System	Ubuntu 18.04
8	Number of CUDA cores	640

C. Use of mobile hardware resources

A single-board special-purpose NVidia Jetson Nano computer uses as a mobile hardware platform for testing. The choice of the platform is justified by the possible further use of this model type as the main algorithm for the operation of the mathematical model of the robot manipulator. TABLE IV shows the characteristics of the test platform.

TABLE IV. NVIDIA JETSON NANO PLATFORM CHARACTERISTICS

#	***Characteristic***	***Value***
1	Processor	Quad-core Cortex-A53 64-bit SoC @ 1.2GHz
2	RAM A storage device	2GB LPDDR4 1600 MHz SDRAM Kingston MicroSDHC 32Gb Class 10 Canvas Select
3	Graphics processor	Nvidia Maxwell architecture with 128 NVidia CUDA Cores
4	CUDA	10.2
5	Operating System	Ubuntu Tegra OS, based on Ubuntu 18.04
6	Versions of Tensorflow, PyTorch	Ver. 2.3.1 (TF), ver 1.8 (PyTorch)
7	Python	Ver. 3.8.2 Anaconda x64
8	Number of CUDA cores	128

IV. Model Optimizationand Quantization After Learning

Quantization is the finished model optimization by converting it to a data type. It is smaller in terms of the amount of memory occupied and, accordingly, in speed. This optimization technique makes it possible to reduce the amount of memory used by the model, as well as to speed up its execution due to the use of other ("lighter") types with minimal loss in terms of accuracy.

$$quantized=real/scale+zero\_point \qquad (1)$$

Where:

real – value before quantization (as a rule, training takes place in single-precision floating-point types);

scale – range scaling factor;

zero_point - the average value of the difference between the range maximum point and the range minimum point.

There are two quantization techniques, depending on the stage at which the process is performing:

- Post-training quantization (PTQ) is a technique that is performed after the model training finalization. It consists of the weight range of the finished model in calculating and compressing it to a smaller type within this range [5, 6, 9].

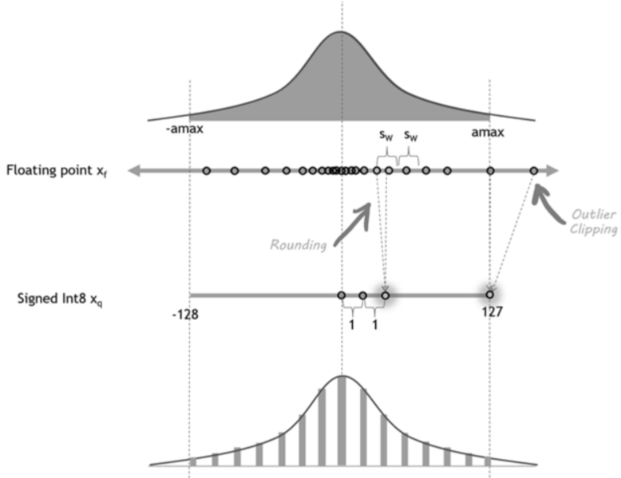


Fig. 2. 8-bit signed integer quantization of float tensor [11].

To increase the accuracy of the transformation (especially on transformations of activation functions) the model is calibrated on a selected representative data set. The advantage of this method is its simplicity and speed. The disadvantage is that this technique significantly degrades the accuracy of the quantized model in some cases.

- Quantization-aware training (QAT) is a technique that allows you to reduce the effect of reducing the accuracy of the model after quantization by adding quantization uncertainty to the total error, which optimizes during training. The advantage of this method is its accuracy, which is significantly higher than in the case of using PTQ. However, this accuracy achieves at the expense of the process.

This study used two different software frameworks with model optimization capabilities. TABLE V presents the TFLITE capabilities in terms of quantization.

TABLE V. Quantization Types of the TFLITE Software Framework [4, 7]

#	Quantization types TFLITE		
	Technique	***Benefits***	***Hardware***
1	Dynamic range quantization	4x smaller, 2x-3x speedup	CPU
2	Full integer quantization	4x smaller, 3x+ speedup	CPU, Edge TPU, Microcontrollers
3	Float16 quantization	2x smaller, GPU acceleration	CPU, GPU

Similarly, the TensorRT framework includes the TensorRT Quantization Toolkit, which enables optimization of various operators. Fig. 3 presents TensorRT capabilities in terms of quantization.

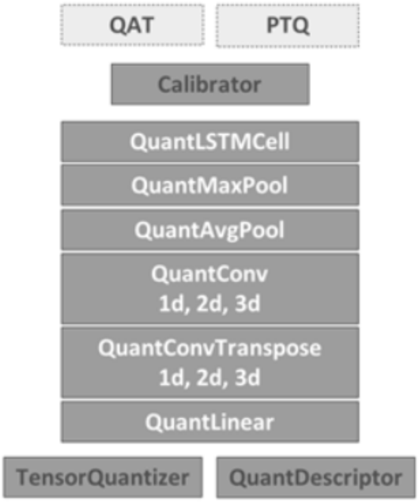


Fig. 3. TensorRT Quantization Toolkit optimization software tool 11].

V. Implementation of the Acceleration Image Recognition on the Mobile Platform NVIDIA Jetson Nano

The NVIDIA Jetson Nano is a small, powerful computer that allows you to run neural networks in parallel for applications such as image classification, object detection. All this in an easy-to-use platform consumes only 5-10 watts. Fig. 4 presents the appearance of the NVIDIA Jetson Nano.

Fig. 4. The appearance of the NVIDIA Jetson Nano.

NVIDIA Jetson Nano equips with an interface for connecting a camera. Photos or videos are recorded on a micro SD Card. It is how images are transferred to a stationary platform. The stationary platform uses for testing the algorithms and the training and validation set of images here.

The main hardware parameters of the NVIDIA Jetson Nano platform are significantly lower compared to similar parameters of a stationary platform based on the seventh-generation Intel Core i5 processor from the point of view of image recognition:

- The number of CUDA cores is five times smaller.
- The speed of the central processor is three times lower.

The NVIDIA Jetson Nano platform's weight, dimensions, and power consumption allow it to use as a mobile platform. The proposed strategy makes use of a monocular imaging recognition system. For photography, a Raspberry Pi Camera ver.2.0 connects to the NVIDIA Jetson Nano. TABLE VI shows the camera characteristics for testing.

TABLE VI. CHARACTERISTICS OF THE RASPBERRY PI CAMERA VER. 2.0.

#	***Parameter***	***Value***
1	Image resolution	5MP Max 2592 x 1944
2	Connection interface	Ribbon Cable
3	Pixel size	1.4 x 1.4 µm
4	Lens	f=3.6 mm, f/2.9
5	Viewing angle	54° x 41°
6	Maximum video resolution	1080p @ 30fps
7	Maximum number of frames per second	480p @ 90fps
8	Video resolution options	1080p @ 30fps, 720p @ 60fps, 480p @ 90fps
9	Sensor size	3.67mm x 2.74mm (1/4" format)
10	Dimensions of the camera module	25mm x 24mm (9mm thickness)

In order to implement the process of pattern recognition acceleration, the data set was formed by using the augmentation of each of the objects. Augmentation was carried out using rotation, affine transformations and scaling in 2D space of each of the objects. Fig. 5shows an example of one the objects augmentation.

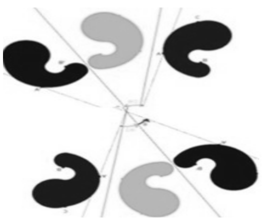


Fig. 5. Augmentation of one of the objects in 2D space.

Both optimized models are launching on the Nvidia Jetson Nano 2 GB mobile platform. The image size for testing is 608x608x3RGB. Fig. 6 shows the data set object types.

Each model's performance was measured on 100 examples. The average execution speed is calculated for each variant. TABLE VII presents the measurement results.

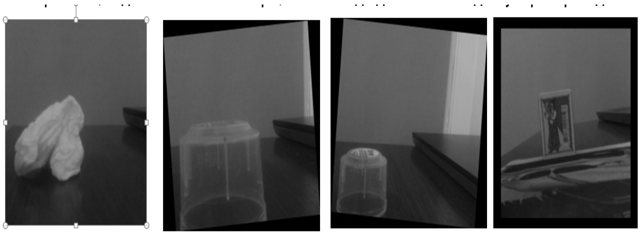


Fig. 6. Types of dataset objects.

TABLE VII. COMPARISION OF THE MODELS PRODUCTIVITY ON THE NVIDIA JETSON NANO PLATFORM

#	***Network***	***Number of para-meters***	***Optimization technology***	***Number of frames per second***	***Average recogni-tion time, ms***
1	YOLOv5s6	3.4 M	Tflite, int8	~10 (tflite)	100
2	YOLOv5s6	3.4 M	N/A	~3 (pt, CPU)	330
3	YOLOv5s6	3.4 M	TensorRT, float32	~12 (pt, CPU)	83
4	YOLOv5s6	3.4 M	TensorRT, int8	~24	42

When testing models trained that have not been optimized to work with the specialized Nvidia Jetson Nano platform, then the results are far from real-time recognition.

The best results were demonstrated by the optimized model for the data types and the Nvidia hardware platform — TensorRT. Photos showed an average recognition time of 42ms. Video showed a speed of 24 frames per second. This can be classified roughly as near-real-time image recognition.

REFERENCES

[1] Nvidia TensorRT Deep Learning documentation. Section 2.4: Types and precision. [Online]. Available: https://docs.nvidia.com/deeplearning/tensorrt/developer-guide/index.html#types-precision

[2] Nvidia TensorRT. [Online]. Available: https://developer.nvidia.com/tensorrt#stories

[3] CUDA Toolkit Documentation. [Online]. Available: https://docs.nvidia.com/cuda/archive/10.2/

[4] TensorFlow Lite guide. [Online]. Available: https://www.tensorflow.org/lite/guide

[5] Post-training quantization. [Online]. Available: https://www.tensorflow.org/lite/performance/post_training_quantization

[6] Post-training quantization. [Online]. Available: https://developer.nvidia.com/blog/achieving-fp32-accuracy-for-int8-inference-using-quantization-aware-training-with-tensorrt/

[7] Integer quantozation for deep learning inference:principles and empirical evaluation. [Online]. Available: https://arxiv.org/pdf/2004.09602.pdf

[8] Scaled-YOLOv4: Scaling Cross Stage Partial Network. [Online]. Available: https://arxiv.org/pdf/2011.08036.pdf

[9] Ian Goodfellow, Yoshua Bengio,Aaron Courville Deep Learning. A MIT Press Book //Ian Goodfellow, Yoshua Bengio,Aaron Courville . - MIT Press, 2016. - 716 p.

[10] Jishu Miao, Tsubasa Hirakawa, Takayoshi Yamashita and Hironobu Fujiyoshi. 3D Object Detection with Normal-map on Point Clouds. [Online]. Available: https://www.researchgate.net/publication/349381918_3D_Object_Detection_with_Normal-map_on_Point_Clouds#pf5

[11] Achieving FP32 Accuracy for INT8 Inference Using Quantization Aware Training with NVIDIA TensorRT. [Online]. Available: https://developer.nvidia.com/blog/achieving-fp32-accuracy-for-int8-inference-using-quantization-aware-training-with-tensorrt/

Machine Learning Assistive State of Charge Estimation of Li-Ion Battery

Saeed Mian Qaisar
CESI LINEACT
Lyon, 69100, France
Electrical and Computer Engineering Department, Effat University,
Jeddah, 22332, Saudi Arabia
smianqaisar@cesi.fr

Ahed Alboody
CESI LINEACT
Nice, 06200, France
aalboody@cesi.fr

Shahad Aldossary, Alhanoof Alhamdan, Nouf Moahammad
Electrical and Computer Engineering Department, Effat University,
Jeddah, 22332, Saudi Arabia

Abdulaziz Turki Almaktoom
Supply Chain Management Department, Effat University,
Jeddah, 22332, Saudi Arabia
abalmaktoom@effatuniversity.edu.sa

***Abstract* — For an effective and economical deployment of battery-powered electric vehicles, mobile phones, laptops, and medical gadgets, the State of Charge (SoC) of the batteries must be properly assessed. It permits a safe operation, have a longer usable battery life, and prevent malfunctions. In this context, the battery management systems provide diverse SoC estimation solutions. However, the Machine Learning (ML) based SoC estimation mechanisms are becoming popular because of their robustness and higher precision. In this study, the features set is prepared using the intended battery cell charge/discharge curves for voltage, current, and temperature. Utilizing statistical analysis and the shape context, the attributes are extracted. Following that, three credible machine learning (ML) algorithms—decision trees, random forests, and linear regression—process the set of mined attributes. The applicability is tested using the Panasonic Lithium-Ion (Li-Ion) battery cells, publicly provided by the McMaster University. The feature extraction and the ML based SoC prediction modules are implemented in MATLAB. The "correlation coefficient", "mean absolute error", and "root mean square error" are used to assess the prediction performance. The results show an outperformance of the random forest regressor among the intended ones by attaining the correlation coefficient value of 0.9988.**

Keywords — Evaluation Measure; Machine learning; MATLAB; Rechargeable Battery; State of Charge

I. Introduction

The new, efficient, environmentally friendly electricity sources are essential. The need to construct effective power systems is driven in part by the desire to preserve the environment and make optimal use of our resources. In order to improve urban air quality and lower pollution levels, bulky lead-acid batteries and oil-guzzling cars could be replaced with high energy density rechargeable batteries that could be used in long-life hybrid plus electric vehicles and renewable hybrid grids [1]. Reliable batteries are consequently required for off-peak electric energy storage.

The deployment of batteries is exponentially exploding. They are almost used in every system such as cell phones to computers, medical gadgets, satellites, and renewable energy based power plants. Batteries are also used as backup power sources in emergencies. For intermittent energy sources, like solar and wind, is another important use [1].

The lithium-Ion (Li-Ion) battery is the most widely used type of rechargeable battery, despite the existence of alternative types [1], [2]. Due to their alluring features, such as being environmentally friendly, having a large discharge depth, having a higher charge-discharge cycle count, having a higher energy density, being compact, having a significantly longer discharge duration, and having a lower maintenance cost, Li-Ion batteries are very popular [3]. Utilizing Li-ion batteries involves a lot of intricacy. It is necessary to inspect the condition of each cell in a battery pack because they might number in the hundreds. As a result, it requires the development of Battery Management Systems (BMSs) of considerable complexity. The event-based techniques can be advantageous in this situation in terms of reducing overhead and improving computing efficiency [4], [5].

BMSs are used more often in contemporary power networks as battery-powered devices like drones, hybrid electric cars, and electric automobiles gain popularity [6]. BMSs are used in these powered systems for a variety of reasons, one of which is that by keeping track of each battery pack cell's health, they make it possible to identify power shortages before they become serious [7]. BMSs are useful for calculating important battery characteristics including the "State of Charge" (SoC) and "Remaining Useful Life" (RUL). The BMS balances the cells, looks for problems, guarantees safety, and regulates the charge-discharge cycles by estimating these parameters.

The SoC calculates the state of charge of a battery cell in respect to its capacity [8]. Different scholars have devised a variety of methods to estimate the SoC [9], [5]. Some of the most popular estimation methods are fuzzy logic, particle filters, kalman filters, impedance spectroscopy, open circuit voltage, and coulomb counting. Accurate real-time SoH estimate is necessary for battery management systems. One of the main benefits of using BMSs in these powered systems is that they help discover power outages early on since they can monitor the health of each battery pack cell. Which of these approaches is most effective depends on how the battery system is used and whether a BMS is required [6], [10], [5].

The Artificial Intelligence (AI) based SoC estimation is becoming popular as compared to the counterparts [12]. It is due to the ever wanted features such as a higher accuracy, adaptability, capability to cope with dynamic conditions,

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

reduced calibration, and scalability. The Machine Learning (ML) based battery SoC estimation method is developed in this regard. In this study, the feature set is prepared using the considered Li-Ion battery cell parameters, such as voltage, current, and temperature. The decision tree, random forest, and linear regression ML methods are then used to process the mined feature set. MATLAB is used to implement the feature extraction and ML-based SoC prediction modules.

The rest of the paper is organized as follow. Section II describes the used materials and methods. Section III presents and described the results and section IV finally concludes the paper.

II. Materials and Methods

The block diagram of suggested method is shown in Fig. 1.

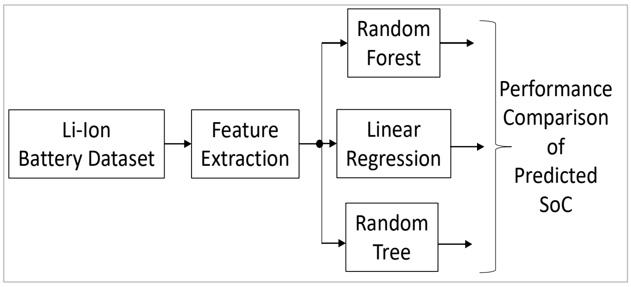


Fig. 1. The block diagram of suggested method.

A. Dataset

The Panasonic Li-Ion battery cells dataset, made available to the public by the universities of Wisconsin-Madison and McMaster, is used to test the applicability [13]. The battery is subjected to a five pulse discharge tests. It is conducted respectively for a range of temperatures and in the following order 25 ºC, 10 ºC, 0 ºC, 10º C, and 20 ºC. The experimentation is performed at certain SoC values: 100%, 95%, 90%, 80%, 70%..., 30%, 25%, 20%, 15%, 10%, 5%, and 0 %. The SoC for each pulse set is approximated using the amp-hour data. In addition to serving as a baseline for HPPC testing, the aforementioned dataset is freely accessible through[1]. It has made significant contributions to the creation of SoC estimate methods and battery models.

B. Feature Extraction

The voltage, temperature, and current charging/discharging curves for battery cells reveal important details about battery life. Additionally, these data from the cell's charging/discharging voltage, temperature, and current can be collected while working separately on the cell's charging and discharging cycles. By tactfully exploring and fusing these features, the current state of the intended Li-ion battery cell can be determined, allowing the employed regression algorithms to determining their SoCs.

In this study, the considered battery cells curves such as the voltage, current and temperature are derived. These curves are derived using their corresponding pulses and the discharges information, provided in the intended dataset. Onward, these pulses are analyzed using the shape context and statistical analysis for mining the pertinent SoC related attributes. The extracted attributes, from consecutive charge/discharge curves of voltage, current, and temperature are fused to form instances [14], [15].

C. SoC Estimation Algorithms

Based on the literature survey three robust SoC estimation algorithms are considered in this study namely, the linear regression (LR), random forest (RF), and random tree (RT). The choice is made on the basis of their frequent use for the battery cells SoC estimation [16]. The performance of considered regression algorithms is evaluated by following the cross validation strategy. In this study the 5-fold cross validation (5-CV) strategy is followed while evaluating the performance of considered regressors. A description of intended regression algorithms is provided in the following.

1) The Linear Regression (LR)

The LR is the most straightforward ML algorithm that can work on complex data patterns. It was built on the premise that the relationship between the input and output variables under study is linear. Thus, this algorithm employs the statistical model that predicts the relationship between the variables based on a linear equation [17]. The Fig. 2 shows an example of the LR algorithm model. This work uses the least square linear regression algorithm. The intercept is set to true and is used in model calculations. It is found that for the studied case the best parameters setting, that provides the highest performance, is attained for the value of complexity parameter that controls the amount of shrinking equal to 0.5.

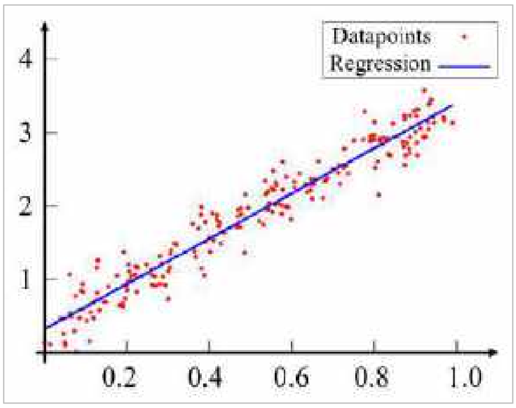


Fig. 2. The LR model principle.

2) The Random Tree (RT)

One of the known supervised ML algorithms is the RT [18]. Its underlying idea is that by combining several weak, randomly produced learners, a strong learner may be created. It generates a large number of decision branches at various nodes; each terminal leaf provides a random classification of the data sample input; the fact that the classifications are random shows that each branch had an equal chance of being sampled. The RT collects the produced decisions and gives a final output with the prediction that got the majority votes [19]. Fig. 3 shows a schematic diagram of the RT algorithm. It shows that the outcome of different prediction branches is combined on the basis of majority voting. The final outcome is the one which secures the highest count of votes. In this work, it is found that the parameter that gives the highest performance is setting the depth of the tree to 15. The algorithm randomly selects the split based on the information gain criterion. For each split, the entropy of each child node is calculated. Onward, the entropy of the split is computed as the

[1] https://data.mendeley.com/datasets/wykht8y7tg/1

weighted average entropy of child nodes. The splitting criteria used to select a split is the highest information gain splitting.

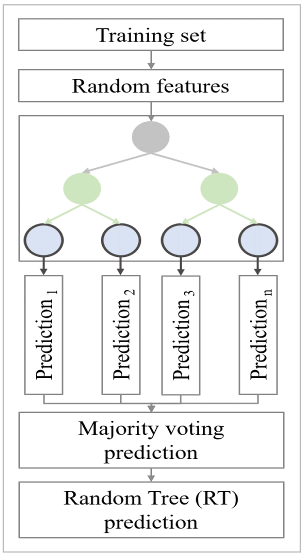


Fig. 3. The RT algorithm's schematic diagram.

3) The Random Forest (RF)

To obtain a more accurate forecast than a single random tree technique, the RF algorithm builds a forest of random trees, or multiple random decision trees at various nodes. The RF gathers these judgments and outputs the result with the highest vote decision after each tree in the RF provides a random classification of the data sample input [19], [20]. Fig. 4 shows a schematic diagram of the RF algorithm. In this work, it is found that the parameters that give the highest performance are setting the number of trees to 50, each with 10 branches. The splitting criteria used is the interaction-curvature. The base learners findings are combined using the majority voting criterion. The outcome is the one which secures the highest count of votes.

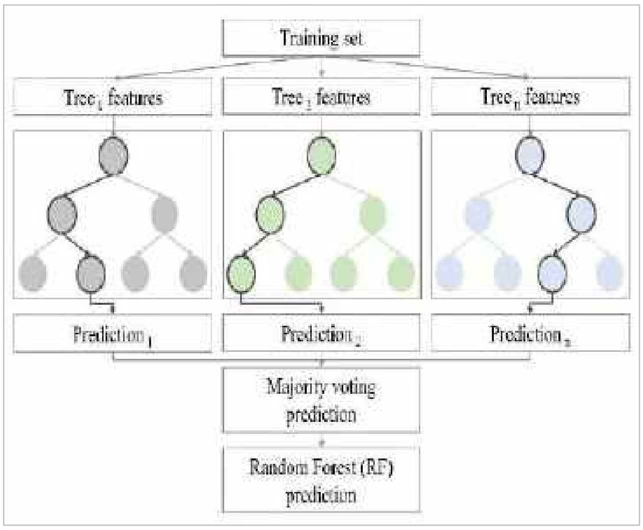


Fig. 4. The RF algorithm's schematic diagram.

D. Evaluation Measure

The "correlation coefficient" (r), "mean absolute error" (MAE), "root mean squared error" (RMSE), "relative absolute error" (RAE), and "root relative squared error" (RRSE) are used to calculate the accuracy of the Li-ion cell SoC forecast under consideration.

The **correlation coefficient $(\boldsymbol{r})$** analysis reveals the degree to which two series of variables are related, such as the actual and expected SoC. It accepts values in the range of -1 to 1. The perfect anti-correlated inverse link is represented by -1, the perfect correlated direct relationship is represented by 0, and the uncorrected series is represented by 0 [21]. The process is given by Eq. (1) [21]. Where, SoC_i^{act} is the actual SoC for the i^{th} instant, SoC_i^{pred} is the predicted SoC, $\overline{SoC^{act}}$ and $\overline{SoC^{pred}}$ are the mean of each one analogously, and n is the total number of the instants.

$$r = \frac{\sum_{i=1}^{n}\left(SoC_i^{act}-\overline{SoC^{act}}\right)\left(SoC_i^{pred}-\overline{SoC^{pred}}\right)}{\sqrt{\sum_{i=1}^{n}\left(SoC_i^{act}-\overline{SoC^{act}}\right)^2 \sum_{i=1}^{n}\left(SoC_i^{pred}-\overline{SoC^{pred}}\right)^2}}. \quad (1)$$

The **mean absolute error (MAE)** and **root mean squared error (RMSE)** are given by Eq. (2) and Eq. (3) respectively [22]. Where, n is the data points or response values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|SoC_i^{act} - SoC_i^{pred}\right|. \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(SoC_i^{act}-SoC_i^{pred}\right)^2}{n}}. \quad (3)$$

The MAE determines the average error whereas the RMSE determines the random component of the data's standard deviation. As a reduced error value is produced, the fit is more beneficial for prediction as their values become closer to 0. The MAE and RMSE can both be used to compare the accuracy of two models if their errors are in similar units.

The **RAE** and **RRSE** are given by Equations (4) and (5), respectively [22].

$$RAE = \frac{\sum_{i=1}^{n}\left|SoC_i^{act}-SoC_i^{pred}\right|}{\sum_{i=1}^{n}\left|SoC_i^{act}-\overline{SoC^{act}}\right|}. \quad (4)$$

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}\left(SoC_i^{act}-SoC_i^{pred}\right)^2}{\sum_{i=1}^{n}\left(SoC_i^{act}-\overline{SoC^{act}}\right)^2}}. \quad (5)$$

The range of the RAE and RRSE is 0 to 1. When their values are lower, the fit is better and the forecast is more accurate the closer they are near 0. They are known as relative mistakes since their computation divides by the SoC_i^{act} variation. They enable model comparisons based on accuracy even if their mistakes are not in the same units, in contrast to the MAE and RMSE, which are meaningless if the error units are different [22].

III. Results and discussion

The performance of devised SoC estimation method is studied by using the Panasonic Li-Ion battery cells dataset. The intended dataset is obtained while performing the five pulse discharge tests. Examples of voltage, current, temperature and power pulses of the intended battery dataset are shown in Fig. 5.

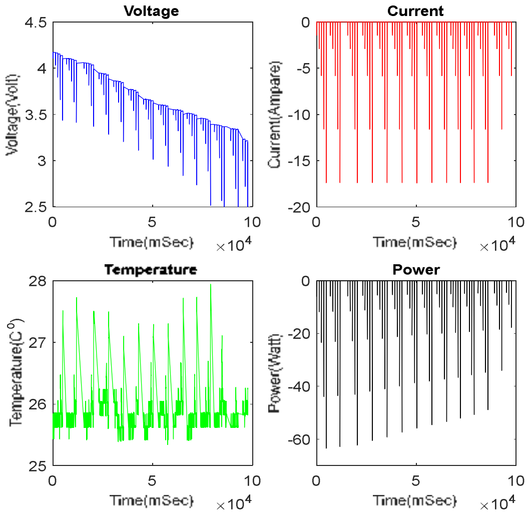


Fig. 5. Examples of pulses of voltage, current, temperature, and power of the intended dataset.

These pulses are used to derive the curves for the battery cells under consideration, including voltage, current, and temperature. Utilizing data from the intended dataset's matching discharges, it is done. After that the features set is prepared as per the shape context and statistical analysis procedure, described in Section II-B.

For an automated SoC prediction, the prepared feature set is processed utilizing the three reliable ML methods. The LR, RF, and RT machine learning algorithms were used in this investigation. Fig. 6 displays the performance in terms of correlation coefficients. Fig. 7 displays the evaluation metrics in terms of the MAE, RMSE, RAE, and RRSE.

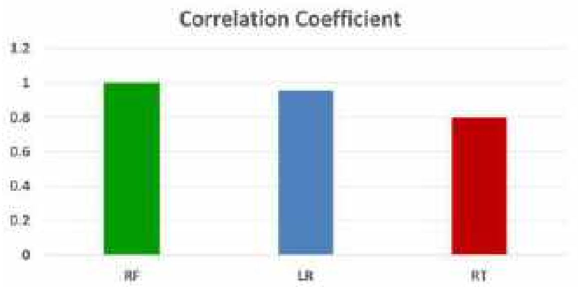


Fig. 6. The correlation coefficient values, obtained respectively with the RF, LR and RT regressors.

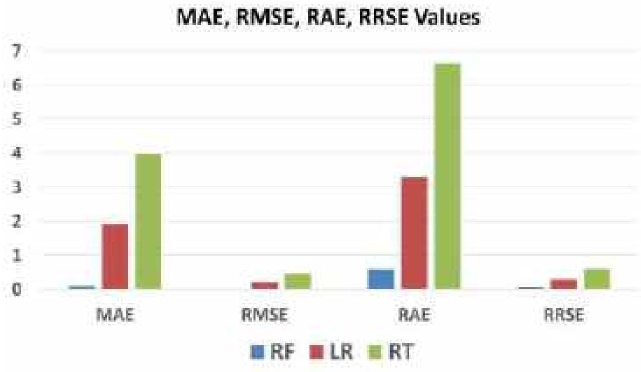


Fig. 7. The MAE, RMSE, RAE, and RRSE values, obtained respectively with the RF, LR and RT regressors.

The summary of findings is presented in Table I.

TABLE I. PERFORMANCE EVALUATION METRICS COMPARISON

Algo.	***r***	**MAE**	**RMSE**	**RAE**	**RRSE**
RF	0.9988	0.0958	0.011004	0.5947	0.053904
LR	0.9546	1.9256	0.22116	3.2864	0.297877
RT	0.7983	3.9759	0.456654	6.6450	0.602293

Fig. 6, Fig. 7 and Table I show that the highest SoC estimation precision is secured by the RF algorithms. It outperforms the LR and RT algorithms in all evaluation measures. It attains the value of correlation coefficient of 0.9988, the MAE score of 0.0958, the RMSE score of 0.011004, the RAE score of 0.5947, and the RRSE score of 0.053904. The second best performer is the LR. It attains the value of correlation coefficient of 0.9546, the MAE score of 1.9256, the RMSE score of 0.22116, the RAE score of 3.2864, and the RRSE score of 0.297877. The RT achieves the least SoC estimation precision among the intended regression algorithms. It attains the value of correlation coefficient of 0.7983, the MAE score of 3.9759, the RMSE score of 0.456654, the RAE score of 6.6450, and the RRSE score of 0.602293.

The RTs often overfit the training data, which means they tend to collect noise and subtleties that may not transfer well to new data. By averaging the predictions of several trees, the RF. This soothes out the peculiarities of each tree and improves generalization.

The LR model may have significant bias or high variance. High variance suggests the model is extremely sensitive to data noise, but high bias suggests the model is too simple to capture underlying patterns. By providing a better bias-variance tradeoff, the RF aids in achieving a balance between these two extremes.

The RF is an ensemble regression technique which mixes several decision trees to produce forecasts. A random portion of the data is used to train each tree, and each tree then generates its own predictions. In contrast to a single decision tree or linear regression, the final prediction is a composite of predictions from all trees, which frequently improves accuracy and generalization. This is the reason of its better performance compared to the LR and RT algorithms.

The integration of event-driven tools and optimization techniques could be beneficial in terms of real-time data processing by enhancing the computational effectiveness and precision [15], [23], [24]. This feasibility could be investigated in future.

CONCLUSION

In this work, a novel hybridization of machine learning and features mining methods based on Li-Ion battery properties is proposed for an automated prediction of the anticipated battery cells' state of charge (SoC). Utilizing the dataset for Panasonic Li-Ion cells, the applicability is examined. The intended dataset's voltage, current, and temperature values are used to construct the battery cells curves. Onward, the generated curves are further analyzed using the shape context and statistical analysis for attributes mining. The mines features from considered voltage, temperature and current curves are fused to prepare instances. The random forest, linear regression and random tree are three robust machine learning regressors used to process the mined features set for an automated prediction of the considered battery cells state of charges. The random forest outperformed the linear regression and random tree and attained the lowest MAE, RMSE, RAE and RRSE values of 0.0958, 0.011004, 0.5947, and 0.053904 respectively. The system also secures the highest correlation coefficient value of 0.9988. The effectiveness of the developed method will be examined in the future for additional prospective datasets. Additionally, the viability of including additional potential deep learning- and machine-learning-based regression models will be examined. Future research will also examine the viability of applying the recommended technique to determine the capacity and health of rechargeable batteries.

References

[1] M. Li, J. Lu, Z. Chen, and K. Amine, "30 years of lithium-ion batteries," Advanced Materials, vol. 30, no. 33, p. 1800561, 2018.

[2] M. H. Amrollahi and S. M. T. Bathaee, "Techno-economic optimization of hybrid photovoltaic/wind generation together with energy storage system in a stand-alone micro-grid subjected to demand response," Applied Energy, vol. 202, pp. 66–77, 2017.

[3] A. K. Thakur et al., "A state of art review and future viewpoint on advance cooling techniques for Lithium–ion battery system of electric vehicles," Journal of Energy Storage, vol. 32, p. 101771, 2020.

[4] S. Mian Qaisar, "Event-driven coulomb counting for effective online approximation of Li-ion battery state of charge," Energies, vol. 13, no. 21, p. 5600, 2020.

[5] S. Mian Qaisar, "A proficient Li-ion battery state of charge estimation based on event-driven processing," Journal of Electrical Engineering & Technology, vol. 15, no. 4, pp. 1871–1877, 2020.

[6] R. Xiong, L. Li, and J. Tian, "Towards a smarter battery management system: A critical review on battery state of health monitoring methods," Journal of Power Sources, vol. 405, pp. 18–29, 2018.

[7] M. A. Roscher, J. Assfalg, and O. S. Bohlen, "Detection of utilizable capacity deterioration in battery systems," IEEE Transactions on vehicular technology, vol. 60, no. 1, pp. 98–103, 2010.

[8] M. H. Lipu et al., "A review of state of health and remaining useful life estimation methods for lithium-ion battery in electric vehicles: Challenges and recommendations," Journal of Cleaner Production, vol. 205, pp. 115–133, 2018.

[9] D. Yang, Y. Wang, R. Pan, R. Chen, and Z. Chen, "State-of-health estimation for the lithium-ion battery based on support vector regression," Applied Energy, vol. 227, pp. 273–283, 2018.

[10] M. A. Roscher, J. Assfalg, and O. S. Bohlen, "Detection of utilizable capacity deterioration in battery systems," IEEE Transactions on vehicular technology, vol. 60, no. 1, pp. 98–103, 2010.

[11] S. M. Qaisar, Electronic management system for rechargeable battery has measuring circuit measuring parameter determining variation of parameter transmitting data to electronic processing unit if variation is higher than predetermined threshold. 2011.

[12] L. Gong, Z. Zhang, Y. Li, X. Li, K. Sun, and P. Tan, "Voltage-stress-based state of charge estimation of pouch lithium-ion batteries using a long short-term memory network," Journal of Energy Storage, vol. 55, p. 105720, 2022.

[13] A. G. Li, A. C. West, and M. Preindl, "Characterizing degradation in lithium-ion batteries with pulsing," Journal of Power Sources, vol. 580, p. 233328, 2023.

[14] L. Ren, L. Zhao, S. Hong, S. Zhao, H. Wang, and L. Zhang, "Remaining Useful Life Prediction for Lithium-Ion Battery: A Deep Learning Approach," IEEE Access, vol. 6, pp. 50587–50598, 2018, DOI: 10.1109/ACCESS.2018.2858856.

[15] S. M. Qaisar and A. E. E. AbdelGawad, "Prediction of the Li-Ion Battery Capacity by Using Event-Driven Acquisition and Machine Learning," presented at the 2021 7th International Conference on Event-Based Control, Communication, and Signal Processing (EBCCSP), IEEE, 2021, pp. 1–6.

[16] G. Dos Reis, C. Strange, M. Yadav, and S. Li, "Lithium-ion battery data and where to find it," Energy and AI, vol. 5, p. 100081, 2021.

[17] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN COMPUT. SCI., vol. 2, no. 3, p. 160, Mar. 2021, DOI: 10.1007/s42979-021-00592-x.

[18] J. Li, W. Ziehm, J. Kimball, R. Landers, and J. Park, "Physical-based training data collection approach for data-driven lithium-ion battery state-of-charge prediction," Energy and AI, vol. 5, p. 100094, 2021.

[19] A. Subasi, Practical machine learning for data analysis using python. Academic Press, 2020.

[20] C. She, Y. Li, C. Zou, T. Wik, Z. Wang, and F. Sun, "Offline and online blended machine learning for lithium-ion battery health state estimation," IEEE Transactions on Transportation Electrification, vol. 8, no. 2, pp. 1604–1618, 2021.

[21] F. Hoffmann, T. Bertram, R. Mikut, M. Reischl, and O. Nelles, "Benchmarking in classification and regression," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 5, p. e1318, 2019.

[22] L. Zavarella, "How to Better Evaluate the Goodness-of-Fit of Regressions," Medium. Accessed: Feb. 24, 2021. [Online]. Available: https://medium.com/microsoftazure/how-to-better-evaluate-the-goodness-of-fit-of-regressions-990dbf1c0091

[23] S. M. Qaisar, "Efficient mobile systems based on adaptive rate signal processing," Computers & Electrical Engineering, vol. 79, p. 106462, 2019.

[24] H. Khan, I. F. Nizami, S. M. Qaisar, A. Waqar, M. Krichen, and A. T. Almaktoom, "Analyzing optimal battery sizing in microgrids based on the feature selection and machine learning approaches," Energies, vol. 15, no. 21, p. 7865, 2022.

Hybradization of Emperical Mode Decomposition and Machine Learning for Categorization of Cardiac Diseases

Danah Milyani
Computer Science Department
Effat University,
Jeddah, 22332, Saudi Arabia

Nouf Mohammad
Electrical and Computer Engineering Department
Effat University,
Jeddah, 22332, Saudi Arabia

Rim Slama
CESI LINEACT,
Lyon, 69100, France

Saeed Mian Qaisar
CESI LINEACT,
Lyon, 69100, France
Electrical and Computer Engineering Department
Effat University,
Jeddah, 22332, Saudi Arabia
smianqaisar@cesi.fr

Alhanoof Alhamdan
Electrical and Computer Engineering Department
Effat University,
Jeddah, 22332, Saudi Arabia

Nora Hamour
CESI LINEACT,
Lyon, 69100, France

***Abstract* — The arrhythmia is one of the cardiovascular diseases which has several types. In literature, researchers have presented a broad study on the strategies utilized for Electrocardiogram (ECG) signal investigation. Automated arrhythmia detection by analyzing the ECG data is reported using a number of intriguing techniques and discoveries. In order to effectively categorize arrhythmia, a novel approach based on the hybridization of the denoising filter, QRS complex segmentation, "Empirical Mode decomposition" (EMD), "Intrinsic Mode Functions" (IMFs) based features extraction, and machine learning techniques is developed in this study. To evaluate the categorization accuracy, the 10-fold cross validation (10-CV) strategy is used. Using an arrhythmia dataset that is publically available for research, the performance of our method is evaluated. A 97% average accuracy score is secured by our method for the problem of 5-class arrhythmias. These findings are comparable or better than counterparts.**

Keywords* — *Arrhythmia; Electrocardiogram; Emperical Mode Decomposition; Features Extraction; Classification; Evaluation Measure; Machine Learning.

I. Introduction

The arrhythmias are of several categories. Mostly, the categorization of arrhythmias is carried out by analyzing the isolated sporadic heartbeats which are derived from the Electrocardiogram (ECG) signals. The arrhythmias mainly cause modification in the morphology or frequency content of the heartbeats and it can be recognized by cardiologists while analyzing the ECG graphs.

For the case of patients with serious cardiac issues an uninterrupted observation is required. In traditional scenario, such patients should be admitted in the hospital which can break their rhythm of life. With the recent development of Internet of Medical Things (IoMT) and cloud processing based mobile healthcare solutions it is possible to realize unobstructive and continuous monitoring of the intended patients cardiac health. It is done by remotely collecting the ECG data via wireless wearables and then analyzing it by cloud applications [1], [2], [3]. For a precise analysis the multichannel ECG recordings are used. It can exponentially raise the volume of data. The manual analysis of such a huge amount of data, at the level of isolated pulses, is a cumbersome task and can result in erroneous diagnosis. In this context, computer based and artificial intelligence (AI) assistive automated solutions are devised [4].

In the aforementioned framework, the two key ECG treatment stages are the pre-processing and isolated heartbeat segmentation [5]. The procedure used in the pre-processing mainly condition the ECG signal and prepare it for the coming stages. The process of segmentation is mainly founded on the basis of QRS complex identification and selection. After pre-processing and obtaining the isolated heartbeats the third and fourth treatment stages are the feature mining and classification. The isolated heartbeats can be further processed in two ways. Firstly, by directly using the deep learning approaches, where the feature mining and classification are embedded as a single stage. In the second approach, the pertinent features are mined from the isolated heartbeats and onward the classification is carried out by processing the mined feature set. In [6]-[18], authors present the key existing techniques for the realization of pre-processing, segmentation, feature extraction, and classification.

In [6], the authors used Wavelet Packet Decomposition (WPD) for the signal denoising and analysis. Onward, entropy based features are mined from sub-bands and the Random Forest (RF) classifier is sued for categorization. In [7], Qaisar et al. used digital filtering for noise removal. The analysis is carried out using the Discrete Wavelet Transform (DWT). The statistical features are extracted from frequency content based selected sub-bands and the classification is done with RF. In [8], Qaisar et al. employed digital filtering for diminishing the noise. The analysis is performed using the Variational Mode Decomposition (VMD). The statistical features are mined form modes. The pertinent features are selected using the Manta Ray Foraging Optimization and the

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

selected feature set is classified using the RF. In [9], authors extracted morphological, time-domain and high order statistical features from the isolated heartbeats. The prepared feature set is processed with the Linear Discriminant (LD) classifier. In [10], authors extracted the morphological and statistical features from the isolated heartbeats. Additionally, the DWT based analysis is performed to mine the sub-band coefficients as extra features. The different features are fused and are onward processed via the Support Vector Machine (SVM) classifier. In [11], several aspects, including the "Local Binary Pattern" (LBP) analysis, DWT, "Higher Order Statistics", and morphological notions are extracted from the isolated heartbeats. The next step is to analyses these attributes, and afterward the most pertinent features are selected using the Manta ray foraging optimization. The SVM classifier then processes the chosen features set. In [12], the deep neural network with supervised contrastive learning and semantic transformations directly processes the isolated heartbeats. In [13], a survey is presented on the computer based automated arrhythmia categorization methods. In [14], an approach is presented for binary class categorization, normal and abnormal, of the isolated heartbeats. The classification is carried out using the SVM classifier. In [15], the VMD is used for the isolated heartbeats analysis. The derived modes are processed to obtain the "Three-dimensional" (3D) "Phase Space Reconstruction" (PSR) together with "Euclidean Distance" (ED) for features mining. The mined feature set is classified by using a deterministic learning approach. In [16], Qaisar has devised an efficient pre-processing approach for the ECG signal conditioning. It is based on the adaptive-rate digital fingering and is particularly appealing for the realization of smart solutions in the context of mobile healthcare. In [18], the DWT is used for feature mining and classification is carried out using the Artificial Neural Network (ANN).

The presented literature review demonstrates the major techniques that may be employed for an automated detection of arrhythmia. In this work, a unique combination of machine learning, intrinsic mode functions (IMFs)-based feature extraction, empirical mode decomposition (EMD)-based analysis, and QRS selection-based segmentation is developed. In order to efficiently process the mined feature set, the effectiveness of two robust k-Nearest Neighbor (k-NN) and Support Vector Machine (SVM) classifiers is compared. The MATLAB program is used to implement the suggested processing modules.

The remainder of this paper is structured as follows. The materials and techniques employed are described in Section II. Section III presents and discusses the findings. Finally, Section IV brings this paper to a conclusion.

II. Materials and Methods

Fig. 1 depicts the suggested system's processing steps. The subsections below provide a description of different modules.

The black colour blocks present the processing stages, already performed on the considered dataset. The blue and golden colour blocks present the processing steps, proposed in this study.

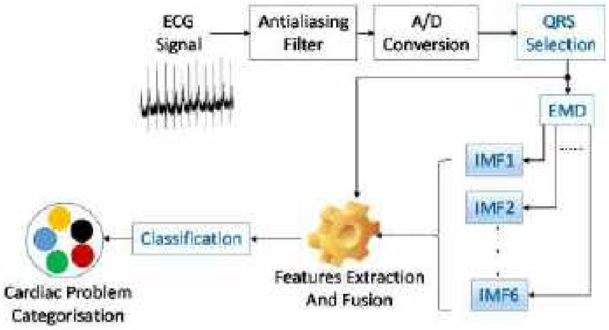


Fig. 1. The proposed system block diagram.

A. Dataset

The arrhythmia center at Boston's Beth Israel Hospital (BIH; currently the Beth Israel Deaconess Medical Center) provided the long-term ECG records that are considered in this article. There are 47 half-hour excerpts of two-channel, from the MIT-BIH arrhythmia database. Twenty-two women and twenty-five men, ranging in age [23, 89] years, are the subjects. For each channel, a digitization rate of 360 samples per second is chosen to properly acquire the ECG data. The data was bandlimited to 60 Hz prior to the digitization.

In this study, the 5 classes of arrhythmia are considered. The class 1 (C1) indicates the "Right Bundle Branch Block" (RBBB) signals. Class 2 (C2) indicates the "Left Bundle Branch Block" (LBBB). Class 3 (C3) indicates "Normal signals" (N). Class 4 (C4) indicates the "Atrial Premature Contraction" (APC). Class 5 (C5) indicates the "Premature Ventricular Contraction" (PVC) signals. For each class, data from three different subjects is collected and 400-instances are included per class. In this manner in total 2000-instances are considered for five intended classes which describes that it is a multi-subject and multi-class categorization problem.

B. QRS Selection

The QRS complex, T wave, and P wave are the three essential elements of a heartbeat (see Fig. 2). For the purpose of identifying an arrhythmia, the QRS complex is crucial [5]. The ECG data was subjected to an amplitude threshold, which enabled the online selection of QRS complexes. The value of threshold is set to match half of the dataset's average R peak amplitudes [5]. Fig. 2 illustrates how the process is done. It shows the magnitude comparison of the incoming samples with the threshold a, which is located on the leading edge of the R pulse. A value for a is chosen that is 50% of the average R peak amplitudes in the dataset under consideration. On the crossover, with respect to a, the i^{th} R peak is detected and the i^{th} QRS complex is segmented, 120 samples around the detected R-peak. It presents a QRS complex with a segmentation length of 300-ms. Based on statistical data on the lengths of the QRS complexes in the target dataset, this segmentation length was chosen. It demonstrates that the targeted dataset's QRS-complex duration is still less than or equal to 250-ms. The process continues and on the detection of next crossover, with respect to a, the $(i+1)^{th}$ R peak is detected and the $(i+1)^{th}$ QRS complex is segmented.

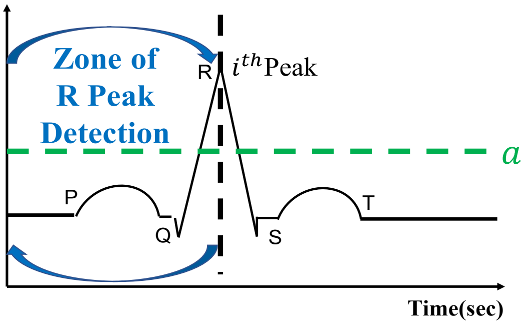


Fig. 2. The Concept behind QRS Selection.

C. Emperical Mode Decomposition (EMD)

A signal is divided up by the EMD into discrete Intrinsic Mode Functions (IMFs) that may be used to analyze the signal [10]. Each IMF acts as the basic oscillatory mode and has an equal number of zero crossings and extrema. Natural signals that are suitable for EMD based analysis are those which display non-stationary and non-linear attributes, such as the ECG. In certain cases, the EMD could leverages the intrinsic features of a signal in a better manner rather than the Fourier Transforms or Wavelet Decomposition. Moreover, the noise cof a signal can be diminished using the EMD and then eliminating the noisy IMFs. However, this process is based on a hypothesis that the noise and the intended signal are unrelated in frequency bands [16].

The upper envelope curve (*xu(t)*) and the lower envelope curve (*xl(t)*) are produced from a signal *x(t)*, respectively, by merging the original signal's maxima and minima points. To compute *G1(t)*, the original signal, *x(t)*, is subtracted from the mean of these two envelopes, M(t) = (xu(t) + xl(t))/2 [10].

$$G_1(t) = x(t) - M(t). \quad (1)$$

This procedure is used to the extraction of the first IMF from *G1(t)* until the requirements for an IMF are met [10]. The first IMF signal, $I_1(t)$, is obtained if the criteria is satisfied. When *x(t)* and $I_1(t)$ are subtracted, the residue, $R_1(t)$, results as:

$$R_1(t) = x(t) - I_1(t). \quad (2)$$

The process may be broadly described as follows for the second, third, and subsequent IMFs:

$$R_{i-1}(t) - I_i(t) = R_1(t); i = 1,2,3,4, \dots \dots, N\,. \quad (3)$$

There are *N* = 6 IMFs in this study as a whole. Finally, six IMFs ($I_1(t)$, $I_2(t)$,..., $I_6(t)$) and a residue signal ($R_N(t)$) are acquired, leading to the following:

$$x(t) = \sum_{i=1}^{N} I_i \ + R_N\ (t). \quad (4)$$

D. Feature Extraction

Each QRS segment's IMFs and its characteristics are identified. The standard deviation of the segmented samples, the minimum absolute value within a segmented sample, the maximum absolute value within a segmented sample, the average of the absolute values of all segmented samples, and the energy of the segmented samples make up the features mined from the QRS segment. The characteristics that may be derived from the extracted IMFs are their standard deviations, minimum and maximum absolute values, mean absolute values of all intended IMFs, ratios between mean absolute values of all IMFs, and energies.

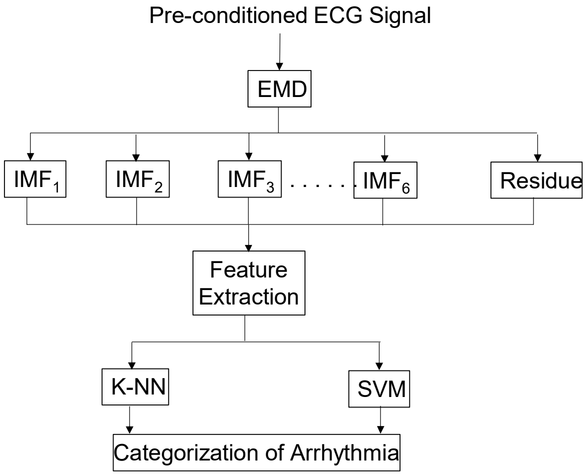


Fig. 3. The EMD process

E. Classification

The robust machine learning-based algorithms "k-Nearest Neighbor" (KNN) and "Support Vector Machine" (SVM) are then used to process the produced feature set. These techniques were chosen because they were often utilized to categorize the ECG signals in earlier investigations. The performance of each classifier is assessed using a 10-fold cross-validation (10-CV) technique and a variety of assessment criteria in order to remove any potential bias in the findings.

The Support Vector Machine (SVM): In contemporary machine learning, SVM (Support Vector Machine) is prized for its resilience and accuracy [19]. Its solid theoretical foundation allows effective operation with small training datasets, regardless of data dimensionality. Moreover, there is a rapid influx of efficient SVM training methods being developed to enhance its practicality.

The SVM seeks to identify the best linear hyperplane that divides two classes in training data in binary classification. In order to categorize fresh data points according to their output sign, this hyperplane is essential. In SVM, it's crucial to maximize the margin, which measures the separation between the nearest data points and the hyperplane [12].

Each data point, symbolized by a feature vector X, represents a singular point within the feature space, which is provided for classification or separation.

$$X \in R^D. \quad (5)$$

For this procedure, where R^D is a vector space with *D* dimensions. Instead of using actual space for *X*, we are employing a comparable notion encompassing the domain, range, and function mapping for the data points. Additionally, this involves further point mapping within the intricate feature space X:

$$\Phi(X) \in R^M. \quad (6)$$

The feature space that results from transforming each input feature and mapping it to a transformed basis vector Φ(x) can be described as:

$$\Phi(X): R^D \rightarrow R^M. \quad (7)$$

The k-Nearest Neighbor (K-NN): The K-Nearest Neighbors (K-NN) algorithm is a classification method that assigns class labels based on the k-closest neighbors, typically using Euclidean distance [19]. It relies on three key components: named items, a distance metric, and the value of k, which sets the number of neighbors.

K-NN faces challenges in selecting an optimal k value, with a small k being sensitive to noise and a large k potentially causing issues with multiple centroids from different classes. Addressing class imbalance is another concern, with a basic compromise approach or a more reliable distance-weighted voting strategy [19].

Efficiently computing k-nearest neighbor distances is crucial for large datasets. Various methods have been developed to reduce computational costs, especially for low-dimensional data, without compromising accuracy [19]. This often involves avoiding the need to calculate distances to all items in the dataset, which can be expensive, particularly for large datasets.

A majority vote from the observation's neighbors is considered when classifying a new observation. Then, using a distance function (as shown in Equation 8), the observation is assigned to the class that is most common among its k closest neighbors. The observation is placed in the class of its nearest neighbor if k is equal to 1.

$$d(x,y) = \sum_{i=1}^{K}(x_i - y_i)^2. \tag{8}$$

F. Cross Validation

In our study, we use k-fold cross-validation, a statistical technique for assessing machine learning models. It helps compare and select the most suitable model for a specific problem. This approach offers several advantages, including simplicity, reduced bias in skill estimation, and ease of implementation It operates by splitting the initial sample into ten subsamples of equivalent size. While the remainder serves as training data, one subsample is utilized for validation [19].

G. Evaluation Measuers

According to [19], each supervised class consists of four different categories. To evaluate the effectiveness of the classifier, a confusion matrix is used. True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) are derived from the confusion matrix [19].

In this study, we consider the following evaluation metrics [19]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \tag{9}$$

$$Precision = \frac{TP}{TP+FP}. \tag{10}$$

$$Recall = \frac{TP}{TP+FN}. \tag{11}$$

$$Specificity = \frac{TN}{TN+FP}. \tag{12}$$

$$F1 - score = \frac{2}{\frac{1}{Recall}+\frac{1}{Precision}}. \tag{13}$$

$$kappa = 1 - \frac{1-p_0}{1-p_e}. \tag{14}$$

$$p_o = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad . \tag{15}$$

$$p_e = \frac{(TP+TN)(TP+FN) + (FP+TN)(FP+FN)}{(TP+TN+FP+FN)^2}. \tag{16}$$

The Receiver Operator Characteristic (ROC): It is a curve demonstrating the relationship between the "True Positive Rate" (TPR) and the "False Positive Rate" (FPR). The Area Under the Curve (AUC): It is also called Area under the ROC curve, which reflects the classifier's ability to differentiate between classes. Higher AUC values indicate improved classifier performance.

III. Results and Discussion

In this part, the classifiers' output from the research is displayed. To determine which categorization technique produces the highest level of prediction, two are evaluated.

TABLE I. THE EVALUATION MEASURES FOR THE K-NN CLASSIFIER

Class	Acc.	Pre.	Rec.	Spec.	F1	KI	AUC
C1	0.96	0.92	0.94	0.98	0.92	0.95	1
C2	0.91	0.83	0.74	0.96	0.78	0.88	0.96
C3	0.92	0.82	0.86	0.95	0.84	0.90	0.98
C4	0.90	0.75	0.81	0.93	0.78	0.86	0.95
C5	0.95	0.88	0.87	0.97	0.88	0.93	0.98
Avg.	0.93	0.84	0.84	0.96	0.84	0.90	0.97

Every class had excellent accuracy results, with the lowest being C4 at 90.00% and the greatest being C1 at 96.00%, as shown by the data in Table I. Additionally, the classifiers' accuracy varied from 75.0% (C4) to 92.0% (C1). The recall value that was attained ranged from 81.00% to 94.00% (C1). The specificity findings are between 98.00% (C1) to 93.00% (C4). Additionally, C1 received the greatest score of 92.00% for F1, while C2 and C4 achieved the lowest score of 78.00%. The kappa values varied between 95.00% (C1) and 86.00% (C4). Lastly, the results from the AUC curve provides 100% for C1 and the lowest AUC value is 95.00% for C4. To sum up all the results, the system secures the best performance for C1 and the lowest for C4.

TABLE II. THE EVALUATION MEASURES FOR THE SVM CLASSIFIER

Class	Acc.	Pre.	Rec.	Spe.	F1	KI	AUC
C1	0.96	0.92	0.90	0.98	0.91	0.95	0.94
C2	0.98	0.94	0.95	0.98	0.94	0.97	0.97
C3	0.95	0.88	0.89	0.98	0.89	0.94	0.93
C4	0.97	0.94	0.92	0.98	0.93	0.96	0.95
C5	0.97	0.93	0.94	0.98	0.94	0.96	0.96
Avg.	0.97	0.92	0.92	0.98	0.92	0.96	0.95

The evaluation measures from SVM classifiers gave better results to that of the KNN classifier (cf. Table II). Here C2 provides the highest values in all the evaluation measures. The least accuracy value of 95.00% is achieved for (C3) and the least precision value of 88.00% is attained for C3. The recall results provides the lowest value of 89.00% for C3. The lowest result for specificity is 98.00% (C4). The least value result from F1 is 89.00% for C3. The kappa statistics result provides the lowest value of 93.00% for C3. The lowest result from the AUC curve evaluation measure is 93.00% for C3. Overall, we can see that that C2 shows the most outstanding results in this classifier.

The SVM secures a better performance by achieving 5.00% superior average accuracy compared to the KNN. It shows that for the studied case the SVM has lesser confusion tendency compared to the KNN. It is mainly attained because of a better pruning capability of the SVM compared to the KNN classifier for the intended application.

The performance of the devised solution is also compared with the previous concurrent approaches. A summary of key findings and methods is presented in Table III. It shows that the devised solution attains a comparable or superior performance.

TABLE III. THE COMPARISON WITH PREVIOUS STUDIES

Study	Features Extraction	Classifier	Acc.
[17]	“Discrete Wavelet Transform” (DWT)	“Probabilistic Neural Network” (PNN)	92.75%
[18]	Wavelet Decomposition + Mutual Information based features selection	“Random Forest" (RF)	97.00%
[20]	“Wavelet Packet Entropy” (WPE)	“Random Forest” (RF)	94.61%
[21]	“Spectrogram images” + “discrete cosine transform (DCT)” based compression	“Convolutional Neural Network” (CNN)	97.00%
[22]	“2D recurrence plot images”	CNN	95.30%
Prop.	EMD + IMFs based features extraction	SVM	97.00%

CONCLUSION

One of the topics with the greatest investigation is the identification of arrhythmias. In order to automatically categorize arrhythmia, a new hybrid method is developed in this paper. It consists a novel hybridization of the QRS selection, empirical mode decomposition, intrinsic mode functions and QRS segments based features extraction, and machine learning technique. The highest classification accuracy of 97.00% is secured for the case of support vector machine classifier while categorizing the five-class arrhythmia dataset. The attained results are promising and encourages the further investigation of this method while incorporating other robust and ensemble learning techniques. In the future, other potential ECG datasets will be used to evaluate the performance of suggested method. Deriving the computational complexity and latency of the suggested method is also a future work. Additionally performance of proposed method will be compared with deep learning techniques.

ACKNOWLEDGMENT

This paper is technically supported by the CESI LINEACT and the Effat University. The authors are thankful to anonymous reviewers for their useful feedback.

REFERENCES

[1] Xia, Y., Zhang, H., Xu, L., Gao, Z., Zhang, H., Liu, H., & Li, S. (2018). An automatic cardiac arrhythmia classification system with wearable electrocardiogram. IEEE Access, 6, 16529-16538.

[2] Chu, J., Wang, H., & Lu, W. (2019). A novel two-lead arrhythmia classification system based on CNN and LSTM. Journal of Mechanics in Medicine and Biology, 19(03), 1950004.

[3] Mian Qaisar, S., & Subasi, A. (2020). Cloud-based ECG monitoring using event-driven ECG acquisition and machine learning techniques. Physical and Engineering Sciences in Medicine, 43, 623-634.

[4] Marinho, L. B., de MM Nascimento, N., Souza, J. W. M., Gurgel, M. V., Reboucas Filho, P. P., & de Albuquerque, V. H. C. (2019). A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification. Future Generation Computer Systems, 97, 564-577.

[5] Qaisar, S. M., Khan, S. I., Dallet, D., Tadeusiewicz, R., & Pławiak, P. (2022). Signal-piloted processing metaheuristic optimization and wavelet decomposition based elucidation of arrhythmia for mobile healthcare. Biocybernetics and Biomedical Engineering, 42(2), 681-694.

[6] Li, T., & Zhou, M. (2016). ECG classification using wavelet packet entropy and random forests. Entropy, 18(8), 285.

[7] Praba, R. A., Suganthi, L., Priya, E. S., & Libisha, J. J. (2022, August). Efficient Cardiac Arrhythmia Detection Using Machine Learning Algorithms. In Journal of Physics: Conference Series, vol. 2318, no. 1, p. 012011). IOP Publishing.

[8] Qaisar, S. M., Khan, S. I., Srinivasan, K., & Krichen, M. (2022). Arrhythmia classification using multirate processing metaheuristic optimization and variational mode decomposition. Journal of King Saud University-Computer and Information Sciences.

[9] Dias, F. M., Monteiro, H. L., Cabral, T. W., Naji, R., Kuehni, M., & Luz, E. J. D. S. (2021). Arrhythmia classification from single-lead ECG signals using the inter-patient paradigm. Computer Methods and Programs in Biomedicine, 202, 105948.

[10] Singh, R., Rajpal, N., & Mehta, R. (2021). An Empiric Analysis of Wavelet-Based Feature Extraction on Deep Learning and Machine Learning Algorithms for Arrhythmia Classification. International Journal of Interactive Multimedia & Artificial Intelligence, 6(6).

[11] Houssein, E. H., Ibrahim, I. E., Neggaz, N., Hassaballah, M., & Wazery, Y. M. (2021). An efficient ECG arrhythmia classification method based on Manta ray foraging optimization. Expert Systems with Applications, 181, 115131.

[12] Le, D., Truong, S., Brijesh, P., Adjeroh, D., & Le, N. (2023). sCL-ST: Supervised Contrastive Learning with Semantic Transformations for Multiple Lead ECG Arrhythmia Classification. IEEE journal of biomedical and health informatics.

[13] P D, Sai Manoj & Jantsch, Axel & Shafique, Muhammad. (2019). Computer-Aided Arrhythmia Diagnosis with Bio-signal Processing: A Survey of Trends and Techniques. ACM Computing Surveys. 52. 10.1145/3297711.

[14] Li, Peng & Chan, Kap & Fu, Sheng & Krishnan, Shankar. (2006). A Concept Learning-Based Patient-Adaptable Abnormal ECG Beat Detector for Long-Term Monitoring of Heart Patients. Neural Networks in Healthcare: Potential and Challenges. 10.4018/978-1-59140-848-2.ch005.

[15] Zeng, Wei & Yuan, Chengzhi. (2021). ECG arrhythmia classification based on variational mode decomposition, Shannon energy envelope and deterministic learning. International Journal of Machine Learning and Cybernetics. 12. 1-26. 10.1007/s13042-021-01389-3.

[16] Qaisar, S. M. (2020). Cardiogram baseline wander and power line interference elimination by proficient adaptive-rate FIR filtering. Engineering Research Express, 2(2), 025024.

[17] Gnecchi, J. A., Morfin-Magaña, R., Lorias-Espinoza, D., del Carmen Tellez-Anguiano, A., Reyes-Archundia, E., Méndez-Patiño, A., & Castañeda-Miranda, R. (2017). DSP-based arrhythmia classification using wavelet transform and probabilistic neural network. Biomedical Signal Processing and Control, 32, 44–56.

[18] Mian Qaisar, S., & Hussain, S. F. (2023). An effective arrhythmia classification via ECG signal subsampling and mutual information based subbands statistical features selection. Journal of Ambient Intelligence and Humanized Computing, 14(3), 1473-1487.

[19] Subasi, A. (2019). Practical guide for biomedical signals analysis using machine learning techniques: A MATLAB based approach. Academic Press.

[20] Li, T., & Zhou, M. (2016). ECG classification using wavelet packet entropy and random forests. Entropy, 18(8), 285.

[21] Hammad, M., Abd El-Latif, A. A., Hussain, A., Abd El-Samie, F. E., Gupta, B. B., Ugail, H., & Sedik, A. (2022). Deep learning models for arrhythmia detection in IoT healthcare applications. Computers and Electrical Engineering, 100, 108011.

Artificial Intelligence Assistive Fire Detection and Seeing the Invisible Through Smoke Using Hyperspectral and Multi-spectral Images

Ahed Alboody
CESI-LINEACT
Nice, 06200, France
aalboody@cesi.fr

Saeed Mian Qaisar
CESI-LINEACT,
Lyon, 69100, France
Electrical and Computer Engineering Department, Effat University
22332, Jeddah, KSA
smianqaisar@cesi.fr

Gilles Roussel
Laboratoire LISIC, Université du Littoral Côte d'Opale
Calais, 62228, France
gilles.roussel@univ-littoral.fr

Abstract — **The global warming has serious impact on our climate. Due to this, the frequency and the intensity of forest fires is increasing. It has shown serious challenges such as the protection of resources, human and wild life, health, and property. This study focuses on developing an artificial intelligence assistive innovative solution for active fire detection in the context of smart cities and vicinities. This paper addresses spectral analysis, detection and classification of active fires and seeing the invisible through smoke and thin clouds. The appealing applications are in urban surveillance, smart cities, future industries, forests and earth observation. The idea is realizable by using an intelligent hybridization of machine/deep learning models and using multi-sensor images (aerial, satellite). For this purpose, we use hyperspectral images (Visible, Near Infra-red (NIR) and Short-Wave Infrared (SWIR)) from AVIRIS aerial and Multi-Spectral Sentinel-2 satellite images. AVIRIS images are 224 spectral bands of wavelengths with a spatial resolution of 15 meters, which varies from 366nm (nanometers) up to 2500nm. However, AVIRIS image studied for their spectral richness of wavelengths not yet completely exploited by machine and deep learning and in SWIR to detect active fires. While, Sentinel-2 image has 13 spectral bands (Visible, NIR and SWIR) with three spatial resolutions (10, 20 and 60 meters). First, we explain and describe the preparation phase of hyperspectral and multispectral image databases of forest fires. These databases contain hyperspectral and multispectral endmembers data of different sites for forest fires. Then, we conduct a spectral analysis from these endmembers to characterize the hyperspectral/multispectral reflectance of active fires to identify the distinct wavelengths for fire detection. We identify the wavelengths that can be used for an effective identification of fire and to see through fires smoke and thin clouds. Onward, the selected feature set is processed by robust machine/deep learning algorithms and their performance is compared for automated identification of fire and invisible vision amelioration. The proposed machine/deep learning method secured an overall test accuracy of 99.1%.**

Keywords* — *Deep learning; Machine learning; Classification; Semantic segmentation; Hyperspectral and Multi-spectral image; Active fire detection; Spectral analysis; Earth observation; Smart urban surveillance

I. Introduction

In our days, forest fires have been increasing dramatically in fire intensity and frequency in many countries in Europe and Canada. Hyperspectral remote sensing systems were used to detect, identify and characterize forest fires [1, 2]. Hyperspectral systems collect spectral information of wavelengths where the wavelengths vary in spectral range from Visible (400-750 nm) to Near Infra-Red (NIR, 750-1100 nm) and Short-Wave Infra-Red (SWIR, 1100-2500 nm) [2, 3, 4, 5, 6, 7]. Despite being an expensive and complex system, hyperspectral system is robust to detect and identify fires [8, 9]. Indeed, it has been shown that NIR/SWIR hyperspectral systems between 1400nm and 2500nm are promising to identify forest fires based on fire index [8] because the spectral reflectance of active fires have distinct features in this range [6, 7, 8, 9].

Recently, machine learning methods, such as Support Vector Machine (SVM) [7], Artificial Neural Network (ANN) [10], and Random Forests [11, 12] were used to classify images (hyper-spectral [9], multispectral [10, 13, 14] and RGB) in order to detect and classify fires. In [7], authors used hyperspectral images from Hyperspectral PRISMA Italian satellite for fire identification in Australia forests. They explored classification technique based on SVM combined with visual interpretation of PRISMA image for validation as ground truth.

Compared to machine learning models, deep learning models based upon convolutional neural networks (CNN) are proposed to classify fires in [9, 11, 15, 16]. For example, 1-Dimensional Convolution Neural Network (1D-CNN) architectures were developed, trained on hyperspectral PRISMA images to classify wildfires [8, 9]. In [13], authors proposed a Fire-Net deep learning framework. Fire-Net is trained on Landsat-8 multispectral satellite images for the classification of active fires and burned areas. Specifically, three optical spectral bands (Red, Green and Blue) are fused with thermal spectral bands of Landsat-8 images for a more effective detection of active fires.

The main contributions of this paper are the following points: (1) Discussing the advantages of hyperspectral and multispectral images over spectral analysis of Visible/NIR/SWIR spectral bands for hyperspectral active fire detection; (2) Presenting the potential of machine learning models (KNN, SVM, ANN) and deep learning models based on 1-Dimensional Convolution Neural Network to detect and classify active fires. Then the results will be compared and discussed. (3) Discussing the possibility and the benefit to integrate the hyperspectral imaging embedded systems (as similar as to AVIRIS and PRISMA) coupled with machine and deep learning models can open new research opportunities for fire detection in application security of urban surveillance, smart cities, and industrial plants.

The rest of the paper is structured as follows. Section II addresses the description of the study areas, the benchmark datasets of AVIRIS hyperspectral and Sentinel-2 multispectral

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

image, the spectral analysis for active fire detection. In Section III, we will apply a supervised machine and 1D-CNN deep learning models for active fire classification. While in Section IV, the results of the proposed models are presented with a critical discussion. Conclusions are given in Section V.

II. Study Areas, Benchmark Datasets and Spectral Analysis for Active Fire Detection

A. Areas of Interest and Datasets for Active Fire Detection

The first datasets utilized in this paper are Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) hyperspectral images. AVIRIS is an instrument in the real-time of Earth Observation and Remote Sensing. AVIRIS sensor delivers calibrated images of spectral radiance in 224 spectral bands with wavelengths from 366 to 2500 nanometers and a spatial resolution of 15 meters. A hyperspectral image can be represented by a data-cube of two spatial dimensions (rows and colons) and third dimension as a spectral dimension for the number of spectral bands. The Shortwave Infra-red (SWIR) range covering wavelengths from 1400 to 2500 nm, can include significant emitted radiance from fire. The utility of hyperspectral remote sensing images are evaluated for active fire detection [8, 9], and in particular, NIR/SWIR remote sensing images. We used AVIRIS data for the study area of the 2019 California wildfire season that burned across the state of California in US (Fig. 1).

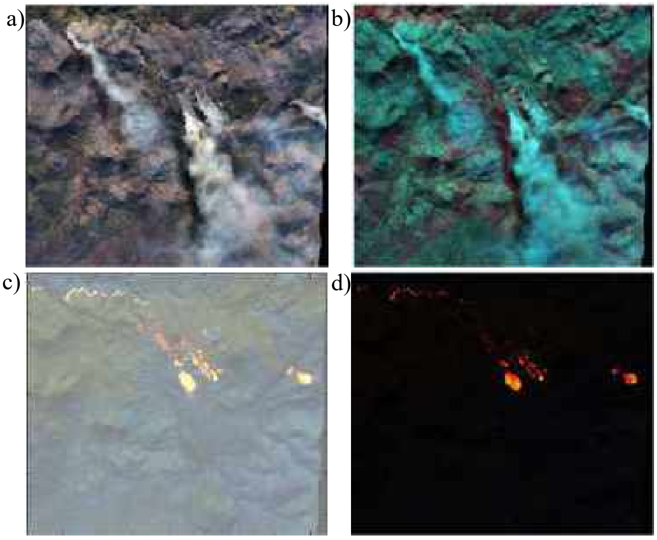


Fig. 1. a) RGB True color composite image from AVIRIS over parts of Sheridan fire in the Prescott National Forest in Arizona, USA on August 21, 2019. This composite used the spectral bands: 647 nm (Red), 550 nm (Green) and 472 nm (Blue); b) color-infrared (CIR) image by dividing the spectral range into three bands : 859nm near infrared (NIR), 647 nm Red, and 550 nm Green bands; c) and d) False color composite inputting active fires monitoring in NIR and SWIR at : 2176 nm (red), 1561nm (green) and 956 nm (blue))

All AVIRIS images are available free on AVIRIS Data Portal[1]. For AVIRIS hyperspectral dataset preparation and the purpose of the accurate evaluation of the detection performance of metrics, we discriminate *between active fire pixels and non-fire pixels (smoke, burned area, vegetation, bare soil, and water)*. To do this discrimination, we select manually pixels with endmembers where each endmember is given by the mean value of spectral reflectance associated with each image patch of size (3×3×224) pixels. The mean value is calculated over all its 9 (3×3) pixels and converted into a vector data of size 1x224 (number of AVIRIS bands). These endmembers are the ground truth that used to train/test the supervised machine and deep learning models for the classification task in Section III. This step of manual selection and classification is needed to determine the ground truth pixels for the implementation of automatic classification based on machine and deep learning. We select these image patches of 9 pixels by exploring the false color composite (Fig. 1c and 1d) and looking at the AVIRIS hyperspectral reflectance, which was comparable with the corresponding classes in [6, 7, 8, 9]. Specifically, *Non-Fires class* contains five subclasses: *smoke, burned areas, vegetation, bare soil and water*. To prepare the ground truth of training datasets, we selected 528 endmembers, which represent pixels of two classes: one for *active fires/Fires class* and the other for *Non-Fires class.* These 528 endmembers are divided into the following number of labeled endmembers: 270 endmembers for active fires, 114 endmembers representing smoke, 18 endmembers for burned areas, 27 endmembers for the bare soil class, 63 endmembers for vegetation, and 36 endmembers for water. We grouped the endmembers of smoke, bare soil, vegetation, burned areas and water into *Non-Fires class*. Finally, *for learning phase*, training/validation datasets have 270 endmembers of active fires for *Fires class* and 258 endmembers for *Non-Fires class* of training data (See Table I). Of the ground truth, we considered a five-fold cross-validation for the training/validation datasets. Then, we selected 106 endmembers for test datasets which are completely different of training/validation datasets. Test datasets are divided into the following number of labeled endmembers: 65 endmembers for active fires, 8 endmembers representing smoke, 13 endmembers for burned areas, 3 endmembers for the bare soil, 6 endmembers for vegetation, and 11 endmembers for water. As similar to training data, test datasets have 65 endmembers of active fires for *Fires class* and 41 endmembers for *Non-Fires class* for *prediction phase*.

The second datasets utilized in this paper are Sentinel-2 multispectral satellite image for Canada and Greece's multiple wildfire in July 2023. Multispectral Imager (MSI) of Sentinel-2 satellite delivers 13 spectral bands with three spatial resolution (10, 20 and 60 meters) [14, 15, 17]. These 13 spectral bands range from the Visible (VNIR) and Near Infra-Red (NIR) to the Short Wave Infra-Red (SWIR). Four spectral bands (Blue (B2), Green (B3), Red (B4), and Near-Infrared (B8)) have a 10-meter spatial resolution and these bands are centered at the following central wavelengths (in nanometers) respectively: 490 nm, 560 nm, 665 nm, 842 nm. Next, six spectral bands in VNIR and SWIR spectral rang are given as follows: red edge (B5, 705nm), near-infrared NIR (B6, 740 nm; B7, 783 nm; and B8A, 865 nm), and short-wave infrared SWIR (B11, 1610 nm; and B12, 2190 nm) which have a 20-meter spatial resolution. Finally, the coastal aerosol (B1, 443 nm), water vapour band (B9, 940 nm), and cirrus (B10, 1375 nm) spectral bands have a 60-meter spatial resolution. Where for the correction of atmospheric effects (e.g., aerosols, cirrus or water vapor), three bands B01, B09 and B10 are used. The remaining ten spectral bands are primarily intended to land use and land cover applications. All Sentinel-2 images are available free on Copernicus Open Access Hub[2]. For the purpose of active fire detection, we use Sentinel-2 level-2A images that are atmospherically corrected surface reflectance in cartographic geometry. At this level, these ten spectral bands, which explored in this study, are B2, B3, B4, B5, B6, B7, B8, B8A, B11 and B12 with spatial resolution of 20-meter

[1] JPL | nasa.gov: https://aviris.jpl.nasa.gov/dataportal/

[2] Copernicus Open Access Hub: https://scihub.copernicus.eu/dhus/#/home

where four bands B2, B3, B4 and B8A are resampled to 20-meter spatial resolution. For the selected area of interest, we obtained the corresponding Sentinel-2 level2A images at the start day or/and 5 days after of the set fire providing *reference data for active fires class*, several days before providing *ground truth for vegetation (of selected area of interest) as non-fires class*, and *after the set fire start date for burned areas as non-fires class.* As similar to AVIRIS visual interpretation in Fig. 1, *false-color images of the SWIR (bands B11 and B12) and NIR (band B8A)* are so useful for visual fire detection (Fig. 2). Similar to AVIRIS datasets, for Sentinel-2 multispectral datasets preparation, Sentinel-2 images are resampled and converted into image patches. Each image is divided into a grid of image patches of size 3×3 pixels. Since, ten VNIR/SWIR spectral bands are considered and the total number of image patches of size (3×3×10) pixels are 1652 manually selected. Then, we select *endmembers* divided into *372 Fires and 384 Non-Fires classes* from these image patches where each endmember of 9 pixels is calculated by the mean multispectral reflectance associated with each image patch of size (3×3×10) pixels. The mean value is calculated over all its 9 (3×3) pixels leading to a vector data of size 1x10 (number of VNIR/SWIR bands of Sentinel-2 images). With supervised machine and deep learning models, we divide the datasets into two parts of *endmembers:* (1) *288 Fires and 284 Non-Fires classes* for training/validation datasets, and (2) the rest for test datasets. The Sentinel-2 multispectral reflectance corresponding to six classes: *active fire, smoke, burned areas, vegetation, bare soil and water* are presented in Fig. 2c.

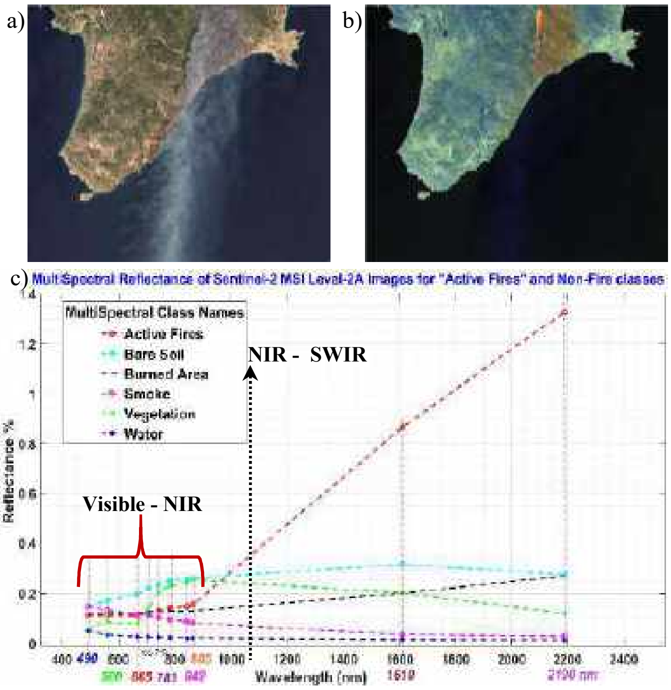


Fig. 2. a) True color composite image from the Sentinel-2 over wildfires burning on the Greek island of Rhodes on July 23, 2023. The RGB composite used the bands centered at 665 nm (B4, Red), 560 nm (B3, Green) and 490 nm (B2, Blue); b) False color composite inputting active fires monitoring in NIR and SWIR at wavelengths 2190 nm (band B12) (red), 1610 nm (band B11) (green) and 865 nm (band B8A). (c) Sentinel-2 Multi-Spectral Reflectance corresponding to six classes: *active fire, bare soil, burned areas, smoke, vegetation, and water.*

B. *Spectral Analysis of Hyperspectral and Multispectral Datacubes for Active Fire Detection*

In this subsection, we perform a spectral analysis of endmembers to characterize the hyperspectral/multispectral reflectance of active fire. Based on the training and test datasets, endmembers were selected by examining the false color composite (Figs. 1c and 1d) and considering the AVIRIS hyperspectral reflectance (see Fig. 3 and Fig. 4). We found spectral discrimination features with specific hyperspectral/multispectral bands in these NIR/SWIR wavelengths: from 1950nm to 2450nm; from 1511nm to 1800nm; from 1166nm to 1332nm; and from 966nm to 1100nm; with very good discrimination features to see through smoke and detect active fires in real-time applications. This spectral analysis shows that the spectral reflectance of fires obtained is similar to that in [7, 8, 9, 10] using PRISMA hyperspectral images. For the detection of active fire using hyperspectral images in industrial environments, we need a NIR/SWIR hyperspectral camera with spectral range from 1100 nm to 2500 nm. For smoke detection in the VNIR range, smoke can be quite easily detected by considering the Visible-NIR bands/wavelengths reported in Fig. 2c and Fig. 3.

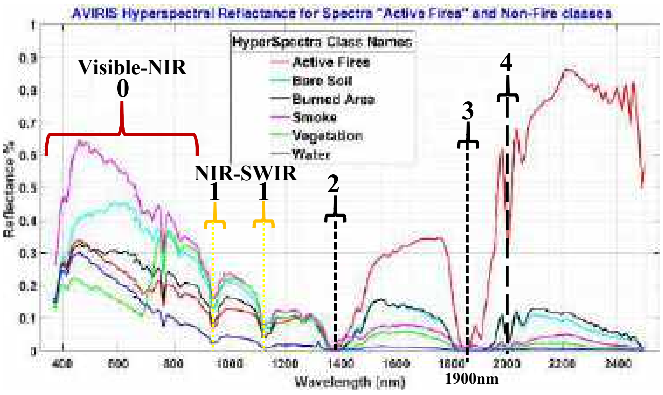


Fig. 3. AVIRIS hyperspectral reflectance corresponding to six classes: *active fire, bare soil, burned areas, smoke, vegetation, and water.* Red bracket number 0 and yellow bracket number 1 indicate the overlapping bands of VNIR-SWIR; Black brackets number 2 and 3 indicate atmospheric attenuation by water vapor, which can occur at 1400 nm and 1900 nm respectively; bracket number 4 indicates the CO_2 absorption bands around 2000 nm.

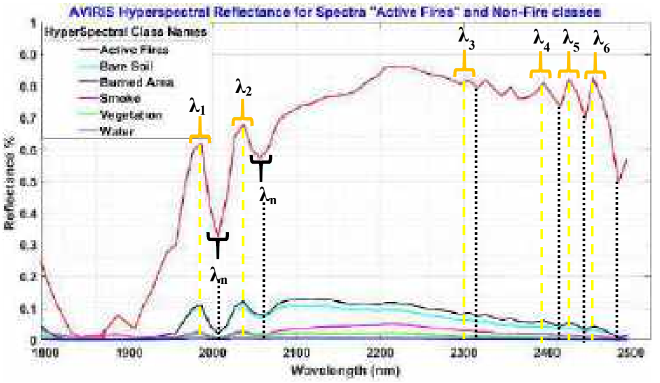


Fig. 4. AVIRIS reflectance for *Fire class* endmember against the background *(Non-Fire classes)* showing CO2 absorption bands around 2000 nm and 2010 nm (λ_n), and around 2050 nm and 2060 nm (λ_n). The locations of the wavelengths corresponding to the "peaks features" of active fires are indicated as λ_1 (absorption features to the left of λ_n) and λ_2 (absorption features to the right of λ_n), respectively.

III. Machine Learning (ML) and CNN Deep Learning (DL) Models for Active Fire Detection

In this section, we will apply a supervised machine and deep learning models to classify active fires using hyperspectral and multispectral images. For this end, we propose to apply three well-known supervised machine learning models (K-Nearest Neighbor classification (KNN), Support Vector Machines (SVM) and Artificial Neural Networks (ANN)) and convolutional deep learning models to classify the endmembers of the training datasets for the learning phase, and the test datasets for the prediction phase, which are described in Section II.A. We describe the proposed supervised machine and deep learning models for active fires detection. For the learning phase in this work, the stopping condition as 30-epochs is defined for all models. The test datasets were used to evaluate the trained models where metrics were calculated in the prediction phase. Finally, we test four trained models with real hyperspectral images to detect active fires.

A. Machine Learning Models (ML)

Based on the training and test datasets, imbalanced classification and weakly supervised learning are challenges for predictive classification because most supervised machine learning models used for classification task were designed around the assumption of an equal number of samples for each class. To take into account these limitations, we have selected three supervised machine-learning models: KNN, SVM and ANN. For each model, there are hyper-parameters to be optimized to determine the best fine-tuning of the classification model by using the training endmembers in *the learning phase* and then to evaluate its performance with the test endmembers *in the prediction phase*. To fine-tune the hyper-parameters of the classification model with challenges of imbalance class and weakly supervised learning, Hyper-Parameter (HP) optimization methods, Bayesian and Random research, offer possibilities to automatically select a classification model with optimized fine-tuning hyper-parameters.

For the issue of fire classification, we therefore propose to apply the main three following phases: *(1) in the first phase, PCA dimensionality reduction by feature extraction* is applied on the training datasets; (*2) in the learning phase*, serval models of supervised machine learning as KNN, SVM and ANN classification models are applied with HP optimization methods of fine-tuning hyper-parameters to determine the best validation accuracy as in [3]. Five-fold cross-validation is considered to protect the trained models against overfitting. This scheme partitions the training datasets into five disjoints fold. Each fold is used once as a validation-fold and the others form a set of training-folds. That allows calculating the size of validation by the 20 percent of the training datasets (Table I).

TABLE I. AVIRIS Datasets for Active Fire Detection

Endmembers extracted from AVIRIS images	**Number of Endmembers for Two classes**		**Total number of endmembers for**
	Fires	***Non-Fires***	
Training endmembers	270	258	528 (83.3 % of Total endmembers)
Validation endmembers (five-fold cross-validation)	54 (20 % of training)	52 (20 % of training)	106 (16.7 % of Total endmembers)
Test endmembers	65	41	106 (16.7 % of Totalendmembers)
Total endmembers	335	299	634 (100%)

For each validation-fold, the classification model is trained using the training-folds and the validation classification accuracy is assessed using the validation-fold. The average validation accuracy is then calculated over all the folds and is used to optimize the fine-tuning parameters of the classification model. These hyper-parameters are determined by an automatic hyper-parameter optimization using two optimization methods: Bayesian and random research. The final validation accuracy have a high estimation of the predictive accuracy of the classification model. (3) *In the prediction phase*, we test the trained models obtained during the learning phase to the test endmembers to determine the overall test accuracy of the trained classification model. Here, the test datasets are completely different of the training datasets. Based on these three phases, Table II presents the top ten-classification models that we tested with different dimensions of the feature subspace obtained by PCA (128, 96, and 64) and with two HP optimization methods. The first column of this Table II gives the name of the tested classification model, the second one indicates the dimension of the features subspace and the third column gives the name of the used HP optimization method. The goal of optimization method is to find a combination of HP values that minimizes an objective function, here the classification error rate. To find this combination, the iteration number of the used method is fixed to 30. The fourth column of Table II describes the determined optimized hyper-parameters. This column is divided into several cells whose number depends on the classification model. For each tested classification model, the validation accuracy computed with the training datasets and the test accuracy computed with the test datasets appear in the fifth and sixth columns, respectively. Accuracy is given as the percentage of endmembers (training or test) that are correctly classified.

B. CNN Deep Learning Model

In this subsection, we propose a lightweight Convolutional Neural Network (CNN) for classification of active fire in hyper/multispectral images that improves the performance of fire detection. The proposed deep learning model based on *a 1-Dimensional Convolution Neural Network (called 1D-CNN)* presented in Fig. 5 and trained on AVIRIS hyperspectral images to classify *active fires*. The *1D-CNN* model includes three convolutional layers, 2 Fully Connected (FC) layers, and one *max-pooling (Max-Pooling)* layer. At the end of the *1D-CNN* model, a *softmax* activation function is applied. This classification model is inspired by the one described in [8, 9]. The input of *1D-CNN model* is the endmember (obtained in Section II.A) comprising the Visible/NIR/SWIR spectral bands of AVIRIS. The input layer of the 1D-CNN model is the *1x224 input features* during the *training/learning phase*. Thus, it is a vector data of size *1xC* where (*C_1=224 is the number of spectral bands for AVIRIS, and C_2=10 for Sentinel-2*). The first layer is one Dimensional (1D) convolutional layer (*Conv1*) with kernel size equal to 1, number of filters equal to *n_1=224*, same padding with *"PaddingValue" equal to "replicate", LeakyReluLayer (value=0.1)* activation function. The second layer is 1D convolutional layer (*Conv2*) with kernel size equal to 3, number of filters equal to *n_2=128*, same padding with *"PaddingValue" equal to "replicate", LeakyReluLayer (value=0.1)* activation function. After these two 1D convolutional layers, a *Max-pooling layer* is connected to the *Conv2* output with these parameters: pooling size of 2 and stride of 2 (with respect Fig. 5, note that *n_3=n_2=128).*The output of *Max-pooling layer* is passed to the third layer of 1D convolutional layer (*Conv3*) with kernel size equal to 3, number of filters equal to *n_2=64*, same padding with *"PaddingValue" equal to "replicate", ReluLayer* activation function. Then, the result of third convolutional layer (*Conv3*) is passed and conncected to the first fully connected layer (*FC1*) of 32 unites with *LeakyReluLayer (value=0.1)* activation function. The last layer is fully connected layer (*FC2*) of 2 unites and *softmax activation function* to classify the output into classes. The 1D-CNN model is trained on *single GPU using Adam optimizer.* The objective function of the 1D-CNN model is *the categorical cross-entropy loss function* of the classification output layer. All hyper-parameters are given in Table II. Models (KNN, SVM, ANN, 1D-CNN) have been implemented using *MATLAB R2021a Machine Learning and Deep Learning Toolbox.*

IV. Results and Discussion

In this section, we present the performance evaluation of active fire detection using the top ten machine and deep learning models. Metrics as confusion matrix for test datasets (Fig. 6) and the validation/test accuracy assessment are given in Table II. The whole results for the proposed models are summarized in Table II with the comparison of the different results. Experimental results, given in Fig. 6, show the test confusion matrix and overall test accuracy for fire classification with four models: (1) Model1-KNN, (2) Model 5-SVM, (3) Model 8-ANN Bi-layered Neural Network, (4) Model 1D-CNN. an overall test accuracy on the test datasets is 99.1% for three different models (Model1-KNN, Model8-ANN, Model 1D-CNN) and 100% for the Model5-SVM, while using the proposed model 1D-CNN achieves high-accuracy to detect active fires on real AVIRIS images as shown as in Fig. 7. Generally, these results are higher than 97.83% where 1D-CNN used in [8, 9] for validation datasets from PRISMA hyperspectral images.

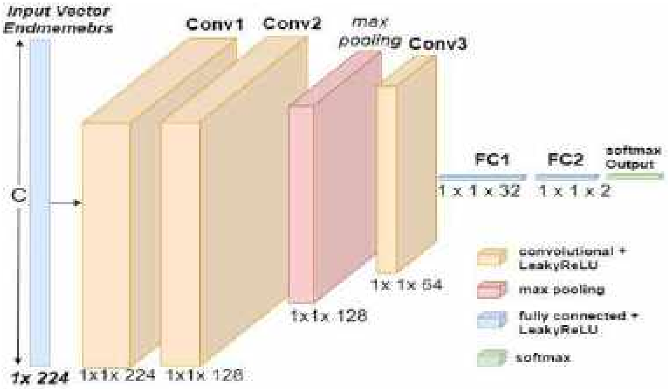


Fig. 5. Architecture of 1D-CNN model for active fires classification of AVIRIS images

TABLE II. Performance Evaluation of Active Fire Detection using The Top Ten Machine and Deep Learning Models

AVIRIS Machine / Deep Learning Models for Active Fire Classification		Value of PCA	Optimizer Method	Fine-tuning / Optimized Hyper-Parameters (HP)				Validation accuracy	Test accuracy
KNN	-			***Number of neighbors***	***Distance metric***	***Distance weight***	***Standardize data***		
	Model 1-KNN	PCA disabled	Bayesian optimization	258	Cosine	Squared inverse	false	99.4%	99.1%
	Model 2-PCA128-KNN	128	Random search	7	Minkowski (cubic)	inverse	false	98.5%	99.1%
	Model 3-PCA96-KNN	96	Random search	1	Euclidean	Inverse	false	98.5%	99.1%
	Model 4-PCA64 -KNN	64	Random search	2	Correlation	Equal	true	98.3%	82.1%
SVM (Kernel scale: 1)	-			***Multi-class method***	***Box constraint level***	***Kernel function***	***Standardize data***		
	Model 5- SVM	PCA disabled	Bayesian optimization	One-vs- All	26.867	Linear	true	99.8%	100.0%
	Model 6-PCA128-SVM	128	Bayesian optimization	One-vs-All	980.8977	Linear	false	95.8%	93.3%
	Model 7-PCA96-SVM	96	Bayesian optimization	One-vs-All	0.0010009	Quadratic	true	97.5%	83.0%
Artificial Neural Network ANN	Model 8-ANN Bi-layered Neural Network	PCA disabled	Bayesian optimization	Number of FC layers: 2	Activation: *Relu*	Regularization strength (Lambda): 5.0497e-08	Standardize data: yes	98.7%	99.1%
				First layer size: 10	Second layer size: 10	Third layer size: 0	Iteration limit: 1000		
	Model 9-PCA128-ANN Tri-layered Neural Network	128	Random search	Number of FC layers: 3	Activation: *ReLU*	Regularization strength (Lambda): 0	Standardize data: yes	90.9%	83.6%
				First layer size: 10	Second layer size: 10	Third layer size: 10	Iteration limit: 1000		
1D-CNN	Model 10-1D-CNN	PCA disabled	Adam Optimizer	Number of FC layers: 2; Convolution layers: three	Number of Max-Pooling layer : one	Activation: *LeakyRelu*/ *ReLU*; *Softmax* for output layer	Regularization strength (l2norm): 1.0e-4	99.4%	99.1%
				First layer size: 32	Second layer size: 2	Mini Batch Size: 12	Learning Rate: 3e-4		

The segmentation map obtained by the prediction phase of the four proposed model are reported in Fig. 7 for the area of interest over parts of Sheridan fire in the Prescott National Forest in Arizona, USA on August 21, 2019. We found that Model 5-SVM and Model 8-ANN model returned false alarms of active fires detection when we tested on real AVIRIS images, while *Model 1D-CNN and Model1-KNN* give best performance for test and real AVIRIS images.

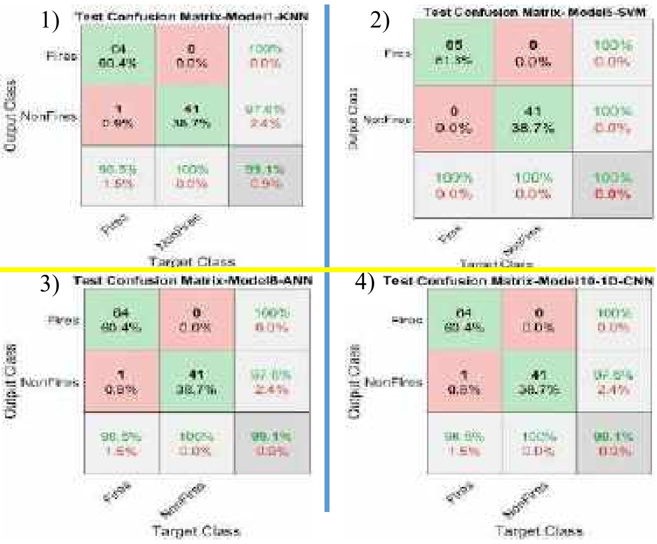


Fig. 6. Test confusion matrix and overall test accuracy for fire classification with: (1) Model1-KNN, (2) Model 5-SVM, (3) Model 8-ANN Bi-layered Neural Network, (4) Model 1D-CNN

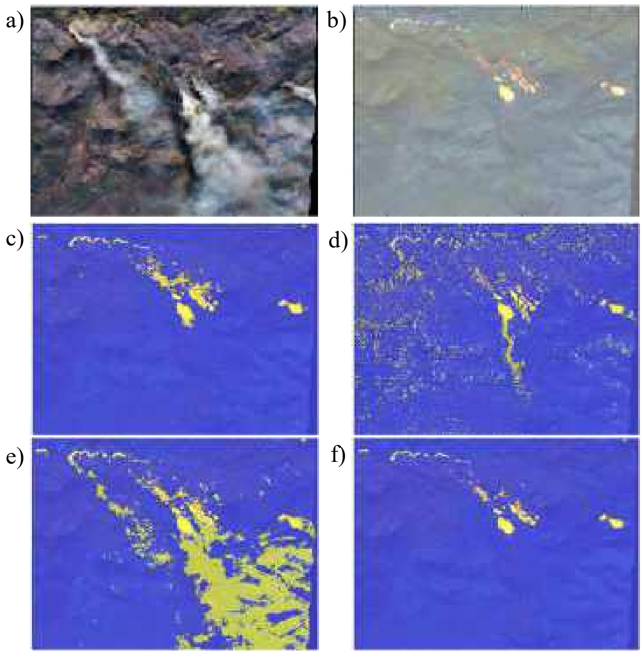


Fig. 7. Segmentation map results (fires in yellow color): (a) RGB AVIRIS aerial, and (b) False-colored images, and the results of the classification prediction from four proposed models: (c) Model1-KNN, (d) Model 5-SVM, (e) Model 8-ANN, and (f) Model 1D-CNN

The results of this paper demonstrates the potentialities of hyperspectral data for active fire detection. The availability of hyperspectral reflectance allows analyzing the hyperspectral information in order to detect and identify fires in smart cities and urban environment. The possibility to use a VNIR/SWIR hyperspectral camera embedded on drone [18, 19] or robot for smart surveillance of fires in urban and industrial environments is one of the bigger advantages of hyperspectral images when we talking about active fire detection in early-warning, real-time smart surveillance and to be considered for future mission dedicated for climate changes and environmental analysis of smart cities.

Conclusion

For active fire detection, this paper has shown the interest of spectral analysis and machine /deep learning coupled with VNIR/SWIR hyperspectral and multispectral images. The results of the proposed models demonstrate that VNIR/SWIR hyperspectral/multispectral images allows to perform many different analysis in order to classify active fires and see through smoke by looking at different spectral bands in NIR/SWIR spectral range. Then, an automatic classification of active fire using supervised machine and deep learning models based on a one-dimensional convolutional layers has been performed. This paper showed also that machine and deep learning models coupled with VNIR/SWIR hyperspectral/multispectral images embedded on an industrial robot or Unmanned Drone is a promising solution to identify and detect active fires in application security of urban surveillance, smart cities, and industrial environments. In future applications, robot and drone's VNIR/SWIR camera embedded may be enable us to easily identify hotspots of active fires.

References

[1] M. K. Griffin, S. M. Hsu, H. H. K. Burke and J. W. Snow, "Characterization and delineation of plumes, clouds and fires in hyperspectral images," In Proc. IEEE International Geoscience and Remote Sensing Symposium '02, vol.2, 2000, pp. 809-812.

[2] P. E. Dennison and D. A. Roberts, "Daytime fire detection using airborne hyperspectral data," Remote Sensing of Environment, vol. 113, pp. 1646-1657, 2009.

[3] A. Alboody, N. Vandenbroucke, A. Porebski, R., F. Viudes, P. Doyen, and R. Amara, "A New Remote Hyperspectral Imaging System Embedded on an Unmanned Aquatic Drone for the Detection and Identification of Floating Plastic Litter Using Machine Learning," Remote Sensing , vol.15, no. 14: 3455, 2023.

[4] S. Veraverbeke, P. Dennison, I. Gitas, G. Hulley, O. Kalashnikova, T. Katagis, L. Kuai, R. Meng, D. Roberts, and N. Stavros, "Hyperspectral remote sensing of fire: State-of-the-art and future perspectives," Remote Sensing of Environment, vol. 216, pp. 105-121, 2018.

[5] R. S. Allison, J. M. Johnston, G. Craig, and S. Jennings, "Airborne Optical and Thermal Remote Sensing for Wildfire Detection and Monitoring," Sensors, vol. 16, no. 8:1310, 2016.

[6] E. Vangi, G. D'Amico, S. Francini, F. Giannetti, B. Lasserre, M. Marchetti, and G. Chirici, "The New Hyperspectral Satellite PRISMA: Imagery for Forest Types Discrimination," Journal of Sensors, vol. 21, no. 4:1182, 2021.

[7] S. Amici, and A. Piscini, "Exploring PRISMA Scene for Fire Detection: Case Study of 2019 Bushfires in Ben Halls Gap National Park, NSW, Australia," Remote Sensing, vol. 13, no 8: 1410, 2021.

[8] D. Spiller, L. Ansalone, S. Amici, A. Piscini, and P. P. Mathieu, "Analaysis and detection of wildfires by using PRISMA hyperspectral imagery," In International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences, XLIII-B3, 215–222, 2021.

[9] K. Thangavel, D. Spiller, R. Sabatini, S. Amici, S. T. Sasidharan, H. Fayek, and P. Marzocca, "Autonomous Satellite Wildfire Detection Using Hyperspectral Imagery and Neural Networks: A Case Study on Australian Wildfire," Remote Sensing, vol. 15, no. 3: 720, 2023.

[10] J. Florath, and S. Keller, "Supervised Machine Learning Approaches on Multispectral Remote Sensing Data for a Combined Detection of Fire and Burned Area," Remote Sensing, vol. 14, no. 3: 657, 2022.

[11] N. Maeda, and H. Tonooka, "Early Stage Forest Fire Detection from Himawari-8 AHI Images Using a Modified MOD14 Algorithm Combined with Machine Learning," Sensors, vol. 23, no. 210, 2023.

[12] K. G. Madhwaraj, V. Asha, A. Vignesh and S. Akshay Shinde, "Forest Fire Detection using Machine Learning," IEEE 12th International Conference on Communication Systems and Network Technologies'23, 2023, pp. 191-196.

[13] S. T. Seydi, V. Saeidi, B. Kalantar, N. Ueda, and A. Abdul Halin, "Fire-Net: A Deep Learning Framework for Active Forest Fire Detection," Journal of Sensors, vol. 2022, no.14, ID 8044390, 2022.

[14] D. A. G. Dell'Aglio, M. Gargiulo, A. Iodice, D. Riccio and G. Ruello, "Active Fire Detection in Multispectral Super-Resolved Sentinel-2 Images by Means of Sam-Based Approach," IEEE 5th International forum on Research and Technology for Society and Industry '19, 2019, pp. 124-127.

[15] M. Gargiulo, D. A. G. Dell' Aglio, A. Iodice, D. Riccio and G. Ruello, "A CNN-based Super-resolution Technique for Active Fire Detection on Sentinel-2 Data," IEEE PhotonIcs & Electromagnetics Research Symposium - Spring (PIERS-Spring), 2019, pp. 418-426.

[16] M. A Akhloufi, R. B. Tokime, and H. Elassady, "Wildland fires detection and segmentation using deep learning. In Pattern recognition and tracking," In Proc. SPIE, vol. 10649, pp.1-12, 2018.

[17] A. Alboody, M. Puigt, G. Roussel, V. Vantrepotte, C. Jamet and T. -K. Tran, "Deepsen3: Deep Multi-Scale Learning Model For Spatial-Spectral Fusion Of Sentinel-2 And Sentinel-3 Remote Sensing Images," In Proc. IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Italy, 2022, pp. 1-5.

[18] M. A. Akhloufi, A. Couturier, and N. A. Castro, "Unmanned Aerial Vehicles for Wildland Fires: Sensing, Perception, Cooperation and Assistance," Drones, vol. 5, no. 1:15, 2021.

[19] Ghali, R.; Akhloufi, M.A.; Mseddi, W.S. "Deep Learning and Transformers Approaches for UAV Based Wildfire Detection and Segmentation," Sensors, vol. 22, no. 5:1977, 2022

Performance Analysis of Database Access: Comparison of Direct Connection, ORM, REST API and GraphQL Approaches

Yaroslav Marchuk
dept. Applied Mathematics
Ivan Franko National University of Lviv
Lviv, Ukraine
Yaroslav.Marchuk@lnu.edu.ua

Ivan Dyyak
dept. Applied Mathematics
Ivan Franko National University of Lviv
Lviv, Ukraine
Ivan.Dyyak@lnu.edu.ua

Ihor Makar
dept. Applied Mathematics
Ivan Franko National University of Lviv
Lviv, Ukraine
Ihor.Makar@lnu.edu.ua

***Abstract* — This study aims to estimate the performance of fetching data by different approaches. Effective access to data is critical for high-performance operations, and understanding the advantages and disadvantages of each system is pivotal for software engineers. The experiments conducted in this study indicated that the REST(REpresentational State Transfer) API(Application Programming Interface)[1] and ORM(Object-Relational Mapping) offered excellent performance, whereas GraphQL[2] required additional resources to achieve comparable efficiency. These results can help businesses choose the applicable system for their specific requirements. The study's findings contribute to the growing knowledge of database access optimization, enabling IT professionals to make informed opinions when designing and enforcing effective database access strategies in their systems. The study also sheds light on the implicit trade-offs between the different database access styles and emphasizes the significance of careful consideration of operating conditions when opting for the most applicable system. All approaches were implemented using the Golang programming language[3] and MongoDB database[4].**

Keywords — REST API, GOLANG, ORM, GraphQL, database.

I. INTRODUCTION

One of the most critical aspects of effective program performance is data accessibility. Data may be accessed in a variety of ways thanks to modern technology. The speed of data access is very important in software development, since slow access can lead to poor performance and longer reaction times.

To work effectively with a database, we must employ an effective data access technique. There are several ways to connect with data, including a database client library, the ORM method, REST API, and GraphQL[5]. Each of these strategies has advantages and disadvantages that must be weighed when selecting the intended outcome for a particular solution.

ORM approach allows to work with the database with a high level of abstraction. The approach to consider is the REST API, which allows you to get and modify data in the database using HTTP requests. Also, a fairly modern approach, GraphQL, has lately surfaced, which provides a more flexible and effective way to fetch and modify data from the database.

In order to ameliorate the performance of database access, it's important to understand how each approach works and what impact they have on the performance of data access. This affects performance and general system effectiveness.

II. ANALYSIS OF RECENT RESEARCH AND PUBLICATIONS

The multitude of studies and publications on database access performance underscore the current relevance of research in this field. To comprehend their benefits and drawbacks, various data access methods have been the subject of multiple experiments and comparisons. High-performance application development mandates a thorough study of all database access methods. Various studies compare programming languages and database access methods while others focus on specific aspects of certain database access methods. Examining these studies is critically important for better understanding the advantages and disadvantages of different methods, and for applying them effectively in practical scenarios. As a result of the analysis, we can conclude that this work complements previous research in this area and opens up new opportunities for building effective software applications.

III. METHODS

We undertake a comparative performance analysis of each of these database access techniques in this study and assess their efficacy in terms of performance and resource usage. We carefully check each database access technique on the same dataset and database server setup in our study. We also consider aspects that might affect performance, such as the frequency of database queries, the number of records, the size of the data, and the presence of indexes.

Object-Relational Mapping(ORM) lets us work with data from the database as with objects of an object-oriented approach. This approach overcomes the problem of incompatibility between object-oriented programming languages and relational databases by introducing an intermediary layer that allows you to interact with the database using an object-oriented approach. ORM enables software engineers to design objects that represent database tables and interact with them in a familiar object-oriented manner. This Object-Relational Mapping eliminates the need for programmers to create SQL queries and manually interface with databases, allowing them to focus on other critical areas of program development. ORM also guarantees data security by checking for valid data input and other database activities automatically.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

The REST API and GraphQL are client-server communication technologies for receiving and transmitting data. To interact with server resources, the REST API uses HTTP methods such as GET, POST, PUT, and DELETE. GraphQL employs a single HTTP endpoint and lets the client request exactly the data it requires, rather than all accessible data, as the REST API does. The REST API frequently delivers full objects that may contain extraneous data. GraphQL, on the other hand, might be more complex to use and configure since you must define the query pattern and the sorts of data that can be requested. In general, the decision between REST API and GraphQL is determined by the project's individual demands and feature requirements.

Golang is a popular programming language for developing high-performance back-end applications that is reliable and fast to execute. Experience shows it is a good choice for creating software that processed enormous volumes of data and interacted with databases. Furthermore, it features a clear syntax and built-in competitiveness support, allowing you to construct efficient multithreaded programs.

IV. Results

This investigation compared four different approaches to database access using Golang: a database access client, the ORM approach, REST API, and GraphQL. The research was conducted on a MongoDB database with 1 500 000 records.

Table 1 and Fig.1-3 shows the results of 4000 requests with 10 connections using each method.

TABLE I. Results of 4000 queries for 10 connections

Method of access	Time of execution, min	Processed requests per second	Average response delay(sec)
Database client	19,93	3,24	1,04
ORM	34,13	1,92	5,46
REST API	24,8	2,58	3,62
GraphQL	39,24	1,59	6,38

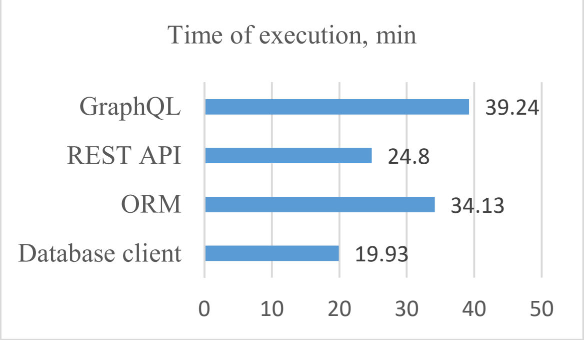


Fig. 1. Duration of complete execution

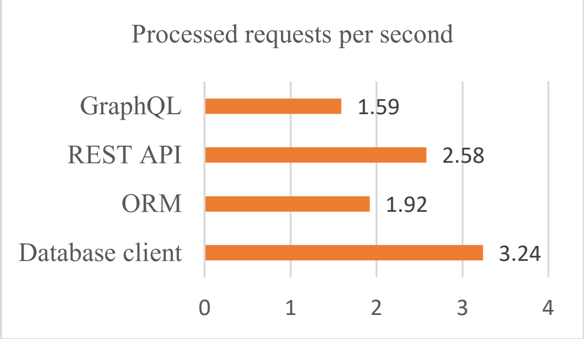


Fig. 2. Requests processed by the server every second

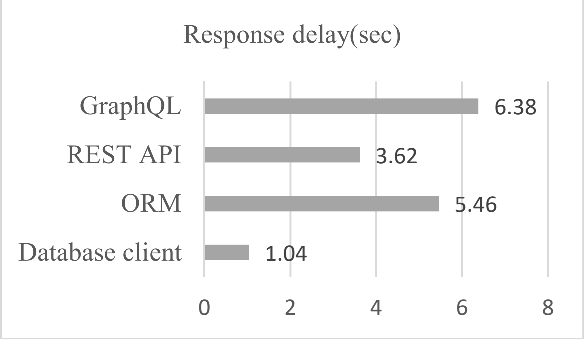


Fig. 3. Average server response delay time

Table 2 and Fig.4-6 shows the data on the execution of 12 000 queries with 40 connections using each method.

TABLE II. Results of 12 000 queries for 40 connections

Method of access	Time of execution, min	Processed requests per second	Average response delay(sec)
Client for database access	40,8	5,08	1,32
ORM	105,2	1,76	12,66
REST API	53,9	3,71	7,1
GraphQL	116,27	1,72	13,02

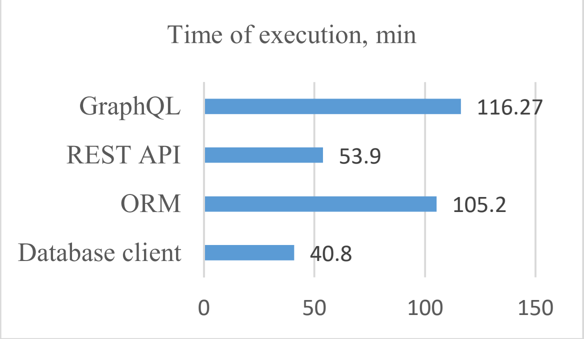


Fig. 4. Duration of complete execution

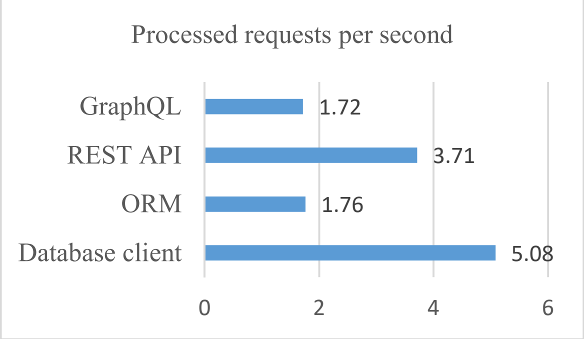


Fig. 5. Requests processed by the server every second

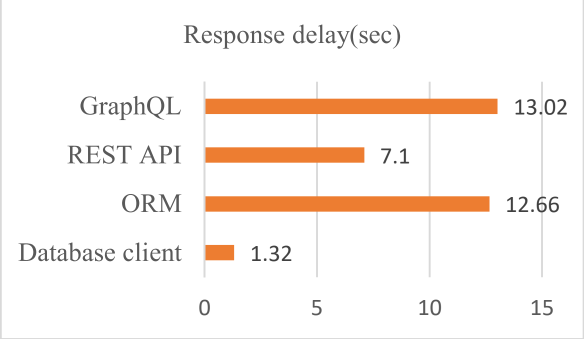


Fig. 6. Average server response delay time

In general, direct access to the database via the client's library was the most productive option. This approach has the quickest query execution time and the shortest response time. It is also worth noting that GraphQL can be handy when working with big volumes of data and different data sources.

According to the numbers in the table, using the MongoDB client to access the database has the highest throughput, while using GraphQL has the lowest. This might be explained by the fact that GraphQL adds an additional layer of abstraction that allows the client to request only the data needed, however this could result in greater time spent on server processing requests. Simultaneously, using the MongoDB client provides direct access to the database and avoids the need for further server processing of queries, resulting in optimal speed.

V. Discussion

In this research, database access performance was examined using various methods such as ORM, REST API, and GraphQL. The study focused on determining the most efficient approach for performance. Software developers working with databases can find the results useful. The results of this research can help them make choices of database access based on their application requirements. The results represent the use of a dataset with limited records and not high-performance hardware. Nevertheless, the research affirmed the significance of selecting the proper method for database access to guarantee optimal application responsiveness. Developers can focus on REST API and ORM approaches that can provide the required performance when working with databases. GraphQL approach would be more effective with a huge amount of data and cloud technology.

VI. Prospects for further research

For further development of the disquisition, a relative analysis of database access speed in cloud technology will be expanded. This will provide a comprehensive understanding of big data processing. Moreover, various cache tools will be used. This can help establish which approaches are more effective for working with different types of databases. We hope that the combination of these technologies will allow us to achieve excellent results.

Conclusions

In our research study, we tested four techniques, each representing a different approach to accessing databases: direct clients, ORM approach, REST APIs, and GraphQL. Following a thorough assessment of the data and information obtained throughout the analysis, the following findings were reached:

The most efficient path towards accessing a database at lightning speed is through implementing the database client technique. Its direct connection grants swift response times and unparalleled throughput capacities. Be warned though that utilizing this approach necessitates investing considerable effort into query construction which may turn out to be challenging in terms of sustainability and maintenance requirements.

The ORM technique simplifies development by allowing you to adopt an object-oriented approach to dealing with data and considerably reducing the amount of code required to write queries. However, as compared to a database client, this method is slightly slower.

The REST API makes it easier to construct client apps by providing a simpler and more native interface for accessing data. However, this method is a little slower than direct database access.

GraphQL provides a more flexible and efficient interface for accessing data, as it allows the client to request exactly the data they need. At the same time, this approach requires more time to process queries compared to other approaches. When choosing GraphQL, you need to optimally configure the components of this solution. It is important to note that an increase in the number of queries to the database results in a decrease in response time and a reduction in the number of queries processed per second. However, the use of multiprocessing and multithreading can enhance the number of requests processed.

References

[1] Naren Yellavula, “Hands-on RESTful Web Services with Go”, Packt Publishing, 2020, 404 p.

[2] E. Porcello, A. Banks, “Learning GraphQL”, O’Reily Media, 2018, 196 p.

[3] Jyotiswarup Raiturkar, “Hands-On Software Architecture with Golang”, Packt Publishing, 2018, 500 p.

[4] A. Phaltankar, J. Ahsan, M. Harrison, and L. Nedov, “MongoDB Fundamentals”, Packt Publishing, 2020, 748 p.

[5] Samer Buna, ”GraphQL in Action”. Manning Publication, 2021, 384 p.

Augmentation in a Binary Text Classification Task

Bohdan Pavlyshenko
Department of electronics and computer technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
bohdan.pavlyshenko@lnu.edu.ua

Mykola Stasiuk
Department of electronics and computer technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
mykola.stasiuk@lnu.edu.ua

Abstract — **The purpose of this paper is to investigate the effect of different kinds of augmentation on the binary text classification performed by different transformer models. Augmentations were performed in three ways: synonym augmentation, contextual word embeddings, and combined. For classification, BERT, ALBERT, DistilBERT, and RoBERTa transformer models were used. It has been found that when using context word embeddings augmentation every model started to overfit during the second training epoch. In the case of combined contextual word embeddings and synonym augmentations utilization, the overfitting issue was overcome, and models exhibited overall good performance. The best performance, however, was obtained when synonym augmentation was used: the overfitting issue was also avoided, and the models' effectiveness was the highest among all experiments.**

Keywords* — *augmentation, text classification, BERT, ALBERT, DistilBERT, RoBERTa.

I. Introduction

Text classification is one of the numerous tasks being solved by Natural Language Processing (NLP). However, for classifiers to be able to process and recognize text correctly, data consistency is one of important aspects. In the modern world, there are 7,168 living languages according to Ethnologue [1] with 4,065 languages currently having a written form. Only a minor part of those languages have enough data to be considered high-resource. These kinds of languages have enough available data, which makes them more attractive for researchers to use in their work. On the other hand, the vast majority of languages are considered low-resource [2]. Moreover, many linguists predict a major part of currently alive languages in the next 100 years [3]. These languages lack available data that can be used in machine learning algorithms. Because of this, a question arises of how to utilize these kinds of languages in machine and deep learning algorithms effectively.

Different approaches were considered, like multilingual projection for parsing low-resource languages [4], and works were performed to apply machine learning techniques without using parallel corpora [5], or utilization of convolutional neural networks and word embeddings [6] and augmentation [7]. Augmentation allows the creation of new data based on already existing, but with different kinds of changes in them, for instance, changing words on their synonyms or using contextual word embeddings. At the same time, augmentation should be used with some caution, as on the one hand it can improve the model's performance and overcome the overfitting issue, but on the other hand, it can introduce new noise to the data, which can worsen the model's effectiveness. However, few researches were done to investigate the effect of these augmentations in connection with transformer models.

II. Methods and Materials

As data for research the dataset [8] was utilized. This dataset contains movie reviews, that have been classified to belong into one of two classes: reviews are positive or negative. The dataset itself consists of two parts - labeled and unlabeled records, but for the experiments only the labeled part was utilized. To have consistent data with nearly balanced records from different classes, the original dataset was shuffled and saved for future augmentation.

When creating datasets for experiments next approach was used: to every record from the original dataset augmentation was applied three times. Thereby after that step in the augmented dataset four records were stored: one from the original dataset and three augmented.

In total four datasets were utilized: an original one, one augmented only with contextual word embeddings, one augmented only with synonyms, and one with combined augmentations. For contextual word embeddings augmentation 'bert-base-uncased' model was used.

In this research, only word-level augmentations were considered. Augmentations were done with the nlpaug library [9]. Two augmentations methods were utilized:

- synonym augmenter - applies semantic meaning based on textual input;

- contextual word embeddings augmenter - applies operation to textual input based on contextual word embeddings.

Augmented and original datasets were used to train four transformer models: BERT [10], DistilBERT [11], RoBERTa [12], and ALBERT [13].

The experiments were conducted on the NVIDIA GeForce GTX 1080 Ti GPU. Due to limited resources, only part of the source dataset was utilized: 20 % or 5000 records from the original. All used models and tokenizers are available on the HuggingFace portal and are accessible by the next names:

- bert-base-uncased;
- distilbert-base-uncased;
- xlm-roberta-base;
- albert-base-v1.

Each model was trained with each dataset in 3 epochs, with a static value of training and evaluation batch size of 8, weight decay was set to 0.01, learning rate was set to $1e^{-5}$, evaluation and save strategies were set to epoch.

To estimate the effect of data augmentations, different transformer models were trained with all four datasets. Evaluation of models' performance was done by different metrics: validation and training losses, accuracy, precession, F1-score, and recall.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

For batch formation, DataColator class from the HuggingFace portal was utilized. Trainer and TrainingArguments classes from the same portal were used for easier feature-complete training.

III. Results and discussion.

From Table 1 we can see the time consumed for datasets preparation with different augmentation. Synonym augmentation was the fastest to complete, and it took nearly 40 times less time than other augmentations. The second fastest was augmentation with contextual word embeddings. This type of augmentation was done with the usage of 'CUDA' for better performance. A dataset with combined augmentations had the longest creational time.

TABLE I. Time of datasets' augmentation

Augmentation	Time, s.ms	
	Train dataset	***Test dataset***
Synonym	184.01	178.00
Contextual word embeddings	7991.01	7783.80
Combined	8313.43	8107.59

TABLE II. Conducted experiments

Experiment name	Used augmentation	Dataset size
Experiment 1	No augmentation used	5000
Experiment 2	Contextual word embeddings with BERT transformer	20000
Experiment 3	Synonym augmentation	20000
Experiment 4	Combined augmentations from experiments 2 and 3	20000

For a clear description of obtained results, names from Table 2 will be used. Hence, the experiment, where the original dataset was utilized will be referred to as experiment 1, experiment 2 will point to the experiment, where a dataset with contextual word embeddings augmentation was used, experiment 3 will mean the experiment, where synonyms augmentation was utilized and experiment 4 will refer to experiment with combined augmentations.

Fig. 1 presents the results of different transformer models training during experiment 1. As can be seen, training losses are decreasing during the learning process, but simultaneously validation losses are increasing. It indicates that models are being overfitted with training data and they can't handle new data in a proper way.

This also can be seen from Fig. 2, as most evaluation metrics showed worse results after the third training epoch than after the first, even though models still had high performance in overall predictions capability, as can be seem from a quite high value of Accuracy metric. From the Precision metric, we can see that even after the second epoch models' capability to correctly produce true positives had decreased, after the third epoch they had only nearly a 10% chance of predicting false positives. Recall shows that after the second epoch the possibility of correctly predicting actual positives rose, after the third epoch it decreased. Finally, F1-score shows that the overall performance of the models is rising after three epochs.

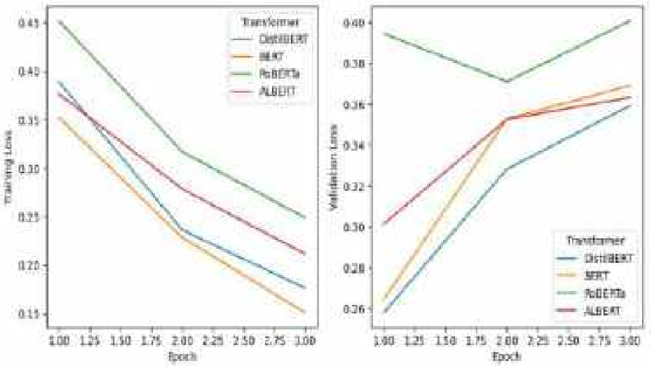


Fig. 1. Training and validation losses of training on the original dataset.

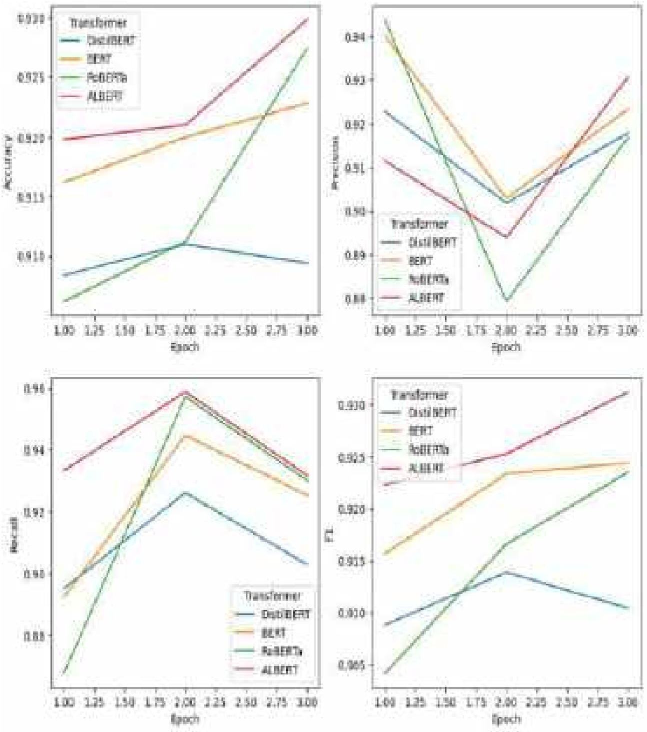


Fig. 2. Evaluation metrics of transformers' performance on the original dataset.

Fig. 3 presents the results obtained during experiment 2. As we can see, this type of augmentation did not affect the overfitting issue. At the same time, transformers' performance got worse from experiment 1, as values of validation losses increased almost twice, in comparison to results obtained when using the original dataset.

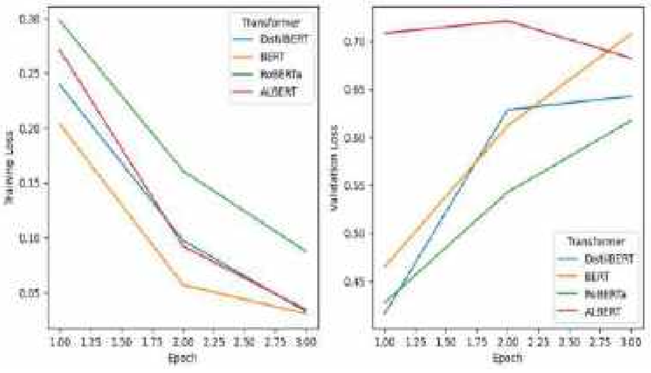


Fig. 3. Evaluation metrics of training on dataset augmented with contextual word embeddings.

At the same time, Fig 4. shows that evaluation metrics also worsened. After three training epochs every transformer, trained during experiment 2, was showing worse results than transformers, from experiment 1.

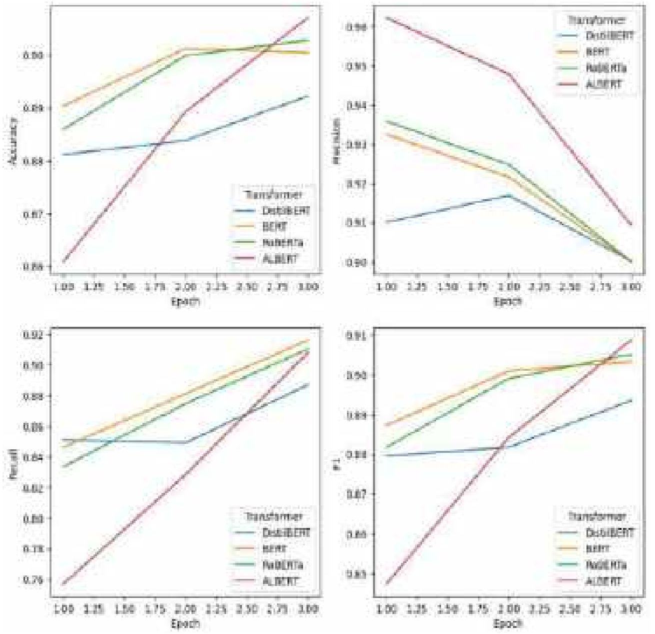


Fig. 4. Evaluation metrics of transformers' performance on dataset augmented with contextual word embeddings.

Fig. 5. demonstrates that transformers, that were trained during experiment 3, exhibit the best performance among all transformers, trained in other experiments. Synonym augmentation helped to overcome overfitting issues, as both, training and validation losses are decreasing dramatically with each learning epoch. Moreover, loss values are the smallest among all experiments.

At the same time, the transformers' performance has also increased for every model, as it is presented in Fig. 6. As we can see, the value of every evaluation metric after the third training epoch is more or equal to 0.995 for every model, which indicates on major growth compared to other experiments. For experiment 3 an additional training epoch can be considered for all transformers, except the ALBERT. ALBERT showed nearly perfect results after three epochs.

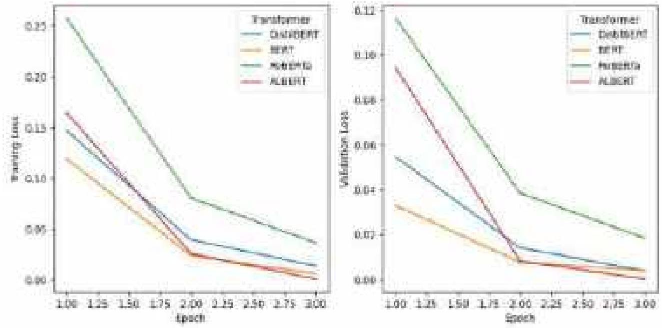


Fig. 5. Training and validation losses of training on the dataset, augmented with synonyms.

As shown in Fig 7. models, trained during experiment 4 were able to solve the overfitting issue. Both training and validation losses were decreasing, with values of losses being greater than the results of models from experiment 3. At the same time, these models showed better performance than models from experiment 1, but worse than models, from experiment 3, as can be seen in Fig. 8.

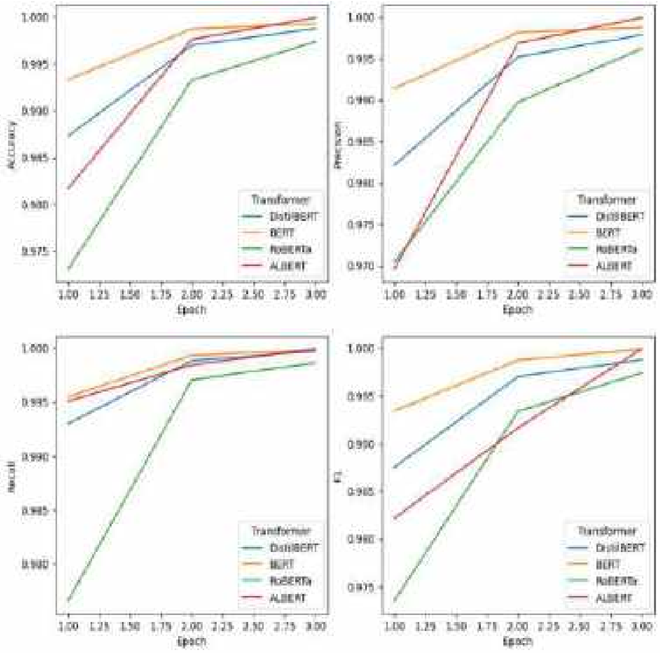


Fig. 6. Training and validation losses of transformers' performance on the dataset, augmented with synonyms.

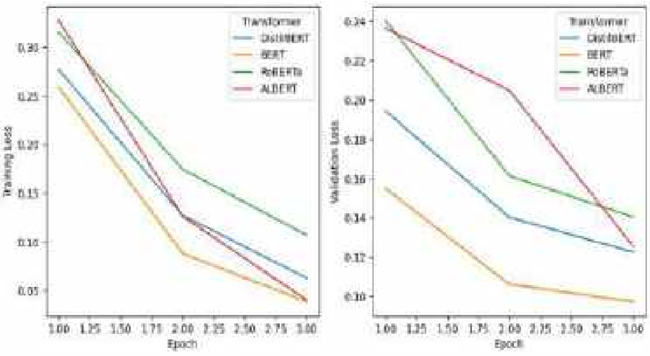


Fig. 7. Training and validation losses of training on the dataset, augmented with combined augmentations.

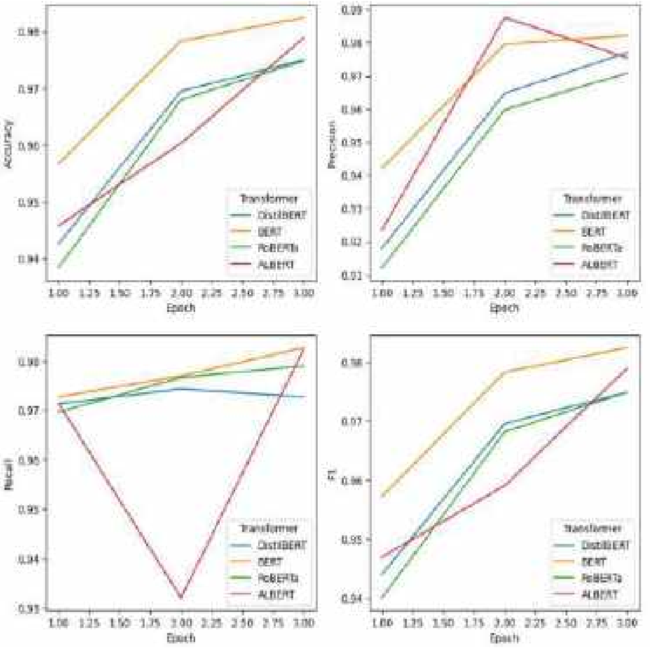


Fig. 8. Training and validation losses of transformers' performance on the dataset, augmented with combined augmentations.

Conclusion

This research paper investigates the effect of word-level augmentation on the performance of different transformer models: BERT, RoBERTa, ALBERT, and DistilBERT. It was found that transformers trained on data augmented only with synonyms augmentation exhibit the best performance among all the models from experiments with other augmented datasets. At the same time, synonym augmenatation alowed the models to solve the overfitting issue. Models, trained on contextual word embeddings augmented dataset showed the worst effectiveness. Models, trained on a combined augmentation dataset, were able to overcome the overfitting issue and show better results, than models trained on the original dataset. In future research, augmentations performed by Large Language Models can be investigated thoroughly to estimate their effectiveness and limitations.

References

[1] D. M. Eberhard, G. F. Simons, and C. D. Fennig (eds.), “Ethnologue: Languages of the World.”, Twenty-sixth edition. Dallas, Texas: SIL International, 2023.

[2] Magueresse A., Vincent C., and Evan H., "Low-resource languages: A review of past work and future challenges.", 2020.

[3] D. Crystal. "Language death. ", Cambridge University Press, 2002.

[4] Ž. Agić, A. Johannsen, B. Plank, H. M. Alonso, N. Schluter, and A. Søgaard, "Multilingual projection for parsing truly low-resource languages.", Transactions of the Association for Computational Linguistics, vol. 4, 2016, pp. 301-312.

[5] A. Karakanta, J. Dehdari, and J. van Genabith, "Neural machine translation for low-resource languages without parallel corpora.", Machine Translation, vol. 32, June 2018, pp. 167-189.

[6] A. Fesseha, S. Xiong, E. D. Emiru, M. Diallo, and A. Dahou, "Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya.", Information, vol. 12(2), 2021, p. 52.

[7] A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales, " Data augmentation for low resource languages.", International Speech Communication Association (ISCA), September 2014, pp. 810-814.

[8] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis." Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, 2011, pp. 142-150.

[9] E. Ma, “NLPAugmentation.”, 2019

[10] J. Devlin, Ming-Wei Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding.", 2018.

[11] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.", 2019.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach.", 2019.

[13] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations.", 2019.

Development of a Virtual Cloud-based Traffic Rules Learning Simulator Using Spherical Video Streams

Artem Kazarian
ACS Department
Lviv Polytechnic National University
Lviv, Ukraine
artem.kazarian@gmail.com

Vasyl Teslyuk
ACS Department
Lviv Polytechnic National University
Lviv, Ukraine
vasyl.m.teslyuk@lpnu.ua

***Abstract* — The article examines the efficiency of spherical video usage for road traffic rule learning and driving skill acquisition in comparison to conventional driving instruction methods. Additionally, it outlines the creation of a virtual educational simulator designed for road traffic rule comprehension and enhancement of driving abilities. The outcomes of this investigation hold significance for road safety and the design of driver education programs. A novel teaching model employing spherical video streaming for road traffic rule understanding has been devised. The research was provided in the scope of Erasmus+ Jean Monnet «Augmented Reality for Education: implementation of European experience» project.**

Keywords — VR, virtual simulator, cloud computing, learning model, spherical video

I. Introduction

Over the past few years, the use of virtual and augmented reality in education and other fields has become an increasingly widespread phenomenon [1, 2, 3]. The utilization of virtual and augmented reality in the learning process creates an open and dynamic experience that engages students and enhances their interest in learning and acquiring new knowledge. Simultaneously, this approach allows learning without the need for specialized equipment and materials, making education more accessible by enabling educational institutions to effectively utilize resources and reduce training costs.

In this study, the effectiveness of using spherical videos for teaching road traffic rules and driving skills is examined, compared to traditional driving instruction methods. Additionally, the implementation of a virtual educational simulator for learning road traffic rules and improving driving skills is described. The results of this research are of significant importance for road safety and the design of driver education programs. The obtained results make it possible to affirm that the developed model works correctly and efficiently for acquiring driving skills. A learning model using spherical video streaming has been developed for the study of road traffic rules.

II. Analysis of Scientific Research

Scientific research on the use of spherical video for teaching road traffic rules and acquiring driving skills has garnered attention in recent years [4]. Some studies have found that spherical video provides a more immersive experience for learners, allowing them to better understand and retain information [5, 6]. This approach can also simulate real-world scenarios and aid in developing driving skills within a controlled and safe environment. However, certain studies also point out technical challenges associated with using spherical video, such as the need for specialized equipment and high-quality cameras, which can increase the cost of organizing the educational process. Furthermore, the lack of standardized rules and recommendations for the use of spherical video in car driving instruction can also impact the effectiveness of learning [7].

III. Development of a Virtual Simulator for Studying Road Traffic Rules

The use of a virtual simulator during the study of road traffic rules offers several significant advantages at the initial stages of learning compared to the traditional practice of driving a real vehicle. The virtual simulator allows experiencing realistic road conditions through spherical videos, ensuring safety during learning without requiring physical presence on the road. The video playback and explanations of depicted road situations enable students to practice in conditions closely resembling real ones and can potentially reduce the number of errors that may occur while driving an actual vehicle, stemming from nervousness associated with the responsibility of driving in real conditions.

Examples of road situations for practice in the developed virtual simulator for teaching road traffic rules include scenarios such as making a U-turn at an intersection, exiting a parking lot, stopping on the road, crossing a railway crossing, driving through an intersection on a green traffic light, and passing through a pedestrian crossing.

To enhance the effectiveness of the educational process, mechanisms have been implemented in which virtual elements in the form of prompts containing information about road traffic rules are superimposed at specified locations on the spherical video. These prompts provide information about the rules of the road applicable in the depicted road situations. When these prompts are displayed, the playback of the video stream is paused to allow the user to familiarize themselves with the provided information and process it more effectively.

The algorithm of operation for the virtual driving skills simulator and road traffic rules learning is depicted in Figure 1. In general, the provided algorithm consists of the next steps:

1) *Loading of spherical video*
2) *Testing or sudy mode selection*
3) *Playing spherical video*
4) *Displaying a questions for knowladge assesment*
5) *Processing and displaying of testing results*
6) *Suggestion to train using another spherical viseos*

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

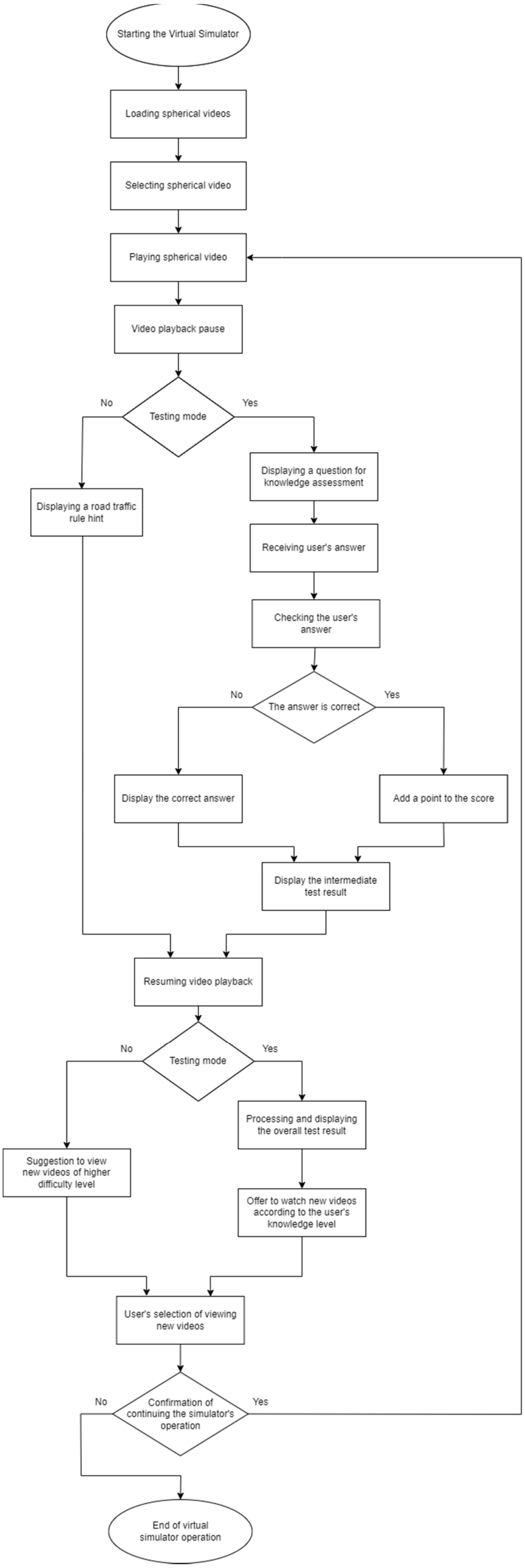

Starting the Virtual Simulator
Loading spherical videos
Selecting spherical video
Playing spherical video
Video playback pause
No
Testing mode
Yes
Displaying a road traffic rule hint
Displaying a question for knowledge assessment
Receiving user's answer
Checking the user's answer
No
The answer is correct
Yes
Display the correct answer
Add a point to the score
Display the intermediate test result
Resuming video playback
No
Testing mode
Yes
Processing and displaying the overall test result
Suggestion to view new videos of higher difficulty level
Offer to watch new videos according to the user's knowledge level
User's selection of viewing new videos
No
Confirmation of continuing the simulator's operation
Yes
End of virtual simulator operation

Fig. 1. The algorithm of operation for the virtual driving skills simulator and road traffic rules learning

User can select one of two modes on the start of simulator work: study or testing mode. In study mode user will see the playing spherical video stream which can be used for studying of right driver actions in different driving situations. In testing mode, the video stream will be stopped in some periods of time and user will see the questions related to current driving situation with four options of answers. After the selection of answer, the user will see the right answer and video stream will be continue played. At the end of lesson user will see the total results of right and wrong answers with total score.

IV. Implementation of the Virtual Simulator

The software implementation of the developed virtual simulator based on a client-server architecture [8, 3], which allows for the distribution of data processing capabilities between the server and the client (web browser). This architecture also enables scalability of the software.

The client-side is implemented as a web application to provide the user interface of the simulator and integrate with virtual reality equipment for playing the spherical video stream. The web application is developed using HTML markup language, cascading style sheets (CSS), and JavaScript libraries.

For video content streams virtual simulator can use spherical video created by any type of 360 VR video camera with support at least 6K resolution which is minimal quality level to provide real-world visualization without textures deformations [10]. Current system realization uses videos recorded by GoPro Max camera.

For implementing the client-side of the virtual simulator, the A-frame library [11] has been chosen. This library is distributed under a permissive license and has a large user base. A-frame offers high-performance capabilities for visualizing 3D scenes and objects, allowing the creation of realistic virtual environments.

The server-side of the virtual simulator is developed using the NodeJS technology [12]. It is responsible for user registration and authentication, storing and streaming the spherical video, managing the user testing logic, and analysing testing results.

When using the virtual simulator, the sense of presence is achieved by allowing users to view a 360-degree video stream around them, as if they are truly inside the car. This feeling of presence encourages more active learning since users feel responsible for decision-making. The sense of presence is created by displaying the spherical image as the user turns their head with a virtual reality headset on. For current research Oculus Quest 2 headset was used. A unique aspect of reproducing spherical video content is that the change in camera angle corresponds to the position of the virtual reality headset at the beginning of the video playback. To create the feeling of presence, this initial camera angle needs to change according to the car's turning angle in the video.

To store the camera rotation data at each specific moment in the video and the data for questions and answer choices that need to be displayed at defined times in the video, a multidimensional time series data structure is used. Each timestamped entry contains information about the car's rotation coordinates at that time or the display coordinates and content of the question block (question text, answer choices, and the correct answer).

The camera's display angle (*visualizationAngle*) is represented as the sum of the car's rotation angle stored in the database for a specified video stream interval and the current rotation angle of the virtual reality headset:

$$visualizatwionAngle = currentAngle + userAngle \quad (1)$$

where *currentAngle* represents the car's rotation angle for the current time of video stream display, and *userAngle* is the current rotation angle of the virtual reality headset.

To prevent excessive data storage in the database, timestamps and their corresponding car rotation angles are stored for time intervals of varying duration, depending on the activity of the car's rotation change (smaller angle changes result in longer intervals between timestamps). To ensure smooth camera rotation display during time periods between database entries, an interpolation method is applied to calculate the camera rotation angle for each second of video stream display.

The first step of interpolation involves searching the database for two adjacent records of car rotation angles ($angle_1$, $angle_2$) with timestamps ($time_1$, $time_2$) between which the current time of video stream display (*currentTime*) falls. The next step is to calculate the duration of the time interval used for interpolation (*timeDiff*), as well as the change in the car's rotation angle during this time interval (*angleDiff*):

$$timeDiff = time_1 - time_2 \quad (2)$$

$$angleDiff = angle_1 - angle_2 \quad (3)$$

Next, we calculate the angle change step for each second within the selected time interval (*angleStep*):

$$angleStep = angleDiff / timeDiff \quad (4)$$

The current time of displaying video content within the interpolation interval (*interpolationTime*) is calculated as the difference between the current time of video content display (*currentTime*) and the timestamp value marking the beginning of the interpolation interval ($time_1$):

$$interpolationTime = currentTime - time_1 \quad (5)$$

The interpolated value of the car's rotation angle at the current time of video content display (*currentAngle*) is calculated using the following formula:

$$currentAngle = angle_1 + interpolationTime * angleStep \quad (6)$$

The obtained interpolated value of the car's rotation angle is used to calculate the camera's display rotation angle (*Formula 1*).

V. Features of User Interface Development

The interface of the developed virtual simulator is depicted in Figures 2-8.

After the successful sign in process user is redirected to the page with available driving lessons. When a user selects one of the lessons the driving lesson page will be opened (Fig. 2.).

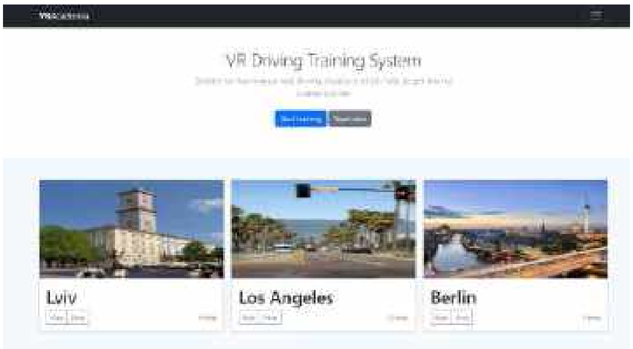


Fig. 2. The screen displays the available driving lessons.

To start training, user need to click the “Start” button. After that the lesson begins.

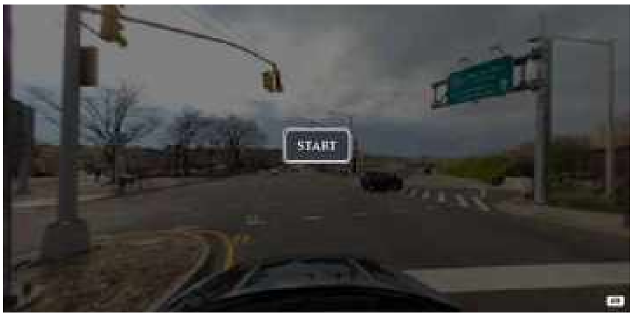


Fig. 3. The screen displays the lesson start screen.

By default, the spherical video stream is visualised in PC screen mode, but users have a possibility to switch into the VR mode and use one of the available VR devices, for example VR headset. Switch into VR mode produced by clicking on “VR” button in the bottom right corner of the screen of PC screen mode.

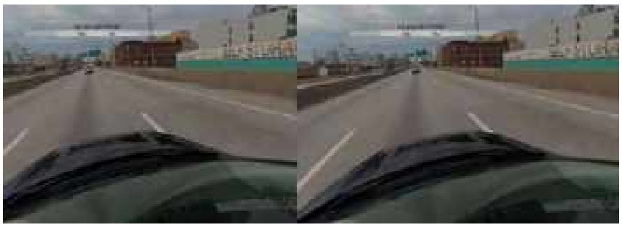


Fig. 4. The screen displays the driving lesson in VR mode.

During the lesson, in defined periods of time the user will see the questions related to the current driving situation with options of answers which the user can select. Each video lesson contains 15 questions related to road situations visualized during the lesson. Average lesson duration is 8-10 minutes. Video stream is paused during the question displaying.

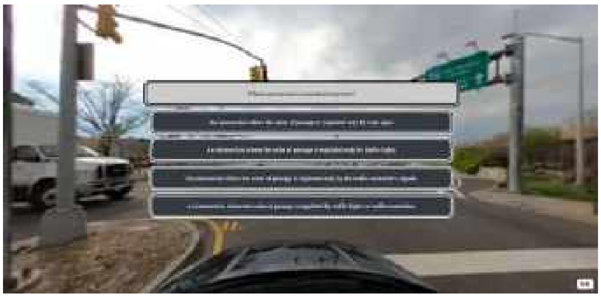


Fig. 5. The screen displays the question related to driving situation.

The selected answer will be highlighted by blue colour.

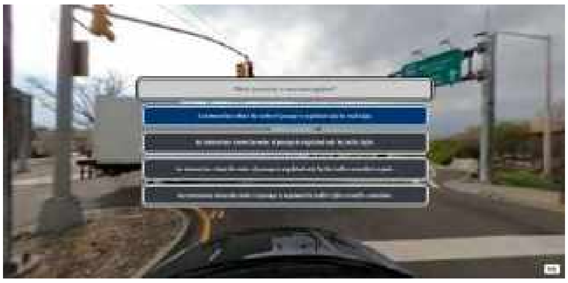


Fig. 6. The screen displays the highlight of user selected question.

After the 3 seconds after the user select answer, the right option will be highlighted by green colour. If the user selects the wrong answer - it will be highlighted by red colour.

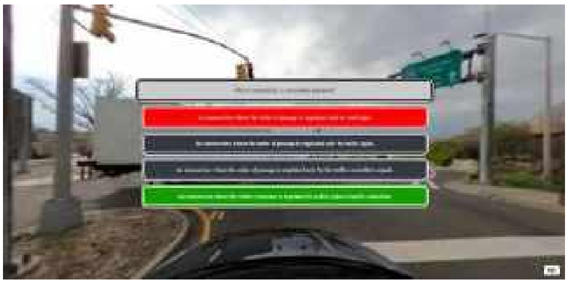
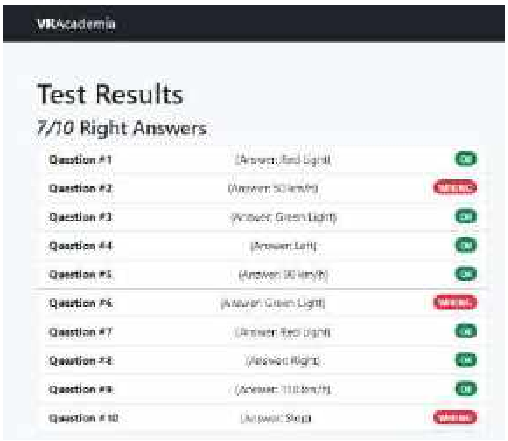


Fig. 7. The screen displays the wrong user selection and right answer.

At the end of the lesson the total result of testing is shown.



VRAcademia

Test Results

7/10 Right Answers

Question #1	(Answer: Red Light)	OK
Question #2	(Answer: 50 km/h)	WRONG
Question #3	(Answer: Green Light)	OK
Question #4	(Answer: Left)	OK
Question #5	(Answer: 30 km/h)	OK
Question #6	(Answer: Green Light)	WRONG
Question #7	(Answer: Red Light)	OK
Question #8	(Answer: Right)	OK
Question #9	(Answer: 110 km/h)	OK
Question #10	(Answer: Stop)	WRONG

Fig. 8. The screen displays the obtained user test results.

Based on the results of conducting training and testing of students at the university department, the developed virtual simulator proves to be an effective tool for learning road traffic rules. After each next training session students gives fewer wrong answers in training mode.

Conclusion

The developed simulator allows users to practice road traffic rules in realistic situations using spherical video. With the help of pop-up hints in specific scenarios, users can choose appropriate actions and receive evaluations of their correctness. The research findings indicate a high level of effectiveness of the simulator in enhancing users' knowledge of road traffic rules.

The integration of spherical videos within VR driving simulations has demonstrated immense potential in revolutionizing driver training, research, and entertainment. By providing a truly immersive experience that closely mimics real-world driving scenarios, this technology offers a safe and controlled environment for drivers to enhance their skills, learn new techniques, and adapt to various challenging situations.

However, challenges remain in optimizing visual quality, reducing motion sickness, and refining the overall user experience. Continued collaboration between VR developers, content creators, and driving experts will be essential to address these hurdles and unlock the full potential of this technology. Additionally, as VR driving simulators become more accessible, considerations of their broader applications, such as autonomous vehicle testing and traffic behavior analysis, come to the forefront.

In summary, the development of VR driving simulators using spherical videos marks a significant stride forward in modernizing driver training and research. As technology continues to advance and our understanding of human-machine interaction deepens, we anticipate a future where VR driving simulators play an integral role in shaping safer, more skilled, and better-prepared drivers.

References

[1] Lege, Ryan & Bonner, Euan. (2020). Virtual reality in education: The promise, progress, and challenge. JALT CALL Journal. 16. 167-180. 10.29140/jaltcall.v16n3.388.

[2] Scavarelli, Anthony & Arya, Ali & Teather, Robert. (2021). Virtual Reality and Augmented Reality in Social Learning Spaces: A Literature Review. Virtual Reality. 25. 10.1007/s10055-020-00444-8.

[3] Lege, Ryan & Bonner, Euan & Frazier, Erin & Pascucci, Luann. (2020). Pedagogical Considerations for Successful Implementation of Virtual Reality in the Language Classroom. 10.4018/978-1-7998-2591-3.ch002.

[4] Lampropoulos, G., Barkoukis, V., Burden, K. et al. 360-degree video in education: An overview and a comparative social media data analysis of the last decade. Smart Learn. Environ. 8, 20 (2021). https://doi.org/10.1186/s40561-021-00165-8.

[5] Barreda-Ángeles, M., Aleix-Guillaume, S., Pereda-Baños, A. (2020). Virtual reality storytelling as a double-edged sword: Immersive presentation of nonfiction 360-video is associated with impaired cognitive information processing. In Communication monographs (pp. 1–20). https://doi.org/10.1080/03637751.2020.1803496.

[6] Berns, A., Mota, J. M., Ruiz-Rube, I., & Dodero, J. M. (2018). Exploring the potential of a 360 video application for foreign language learning. In Proceedings of the sixth international conference on technological ecosystems for enhancing multiculturality (pp. 776–780). https://doi.org/10.1145/3284179.3284309.

[7] Liu, D., Bhagat, K. K., Gao, Y., Chang, T. W., & Huang, R. (2017). The potentials and trends of virtual reality in education. In Virtual, augmented, and mixed realities in education (pp. 105–130). Springer. https://doi.org/10.1007/978-981-10-5490-7_1.

[8] Natesan, Abirami N. (2019). A Detailed Study of Client-Server and its Architecture.

[9] Kazarian A., Teslyuk V., Tykhan M., Mashevska M. Usage of SaaS software delivery model in intelligent house systems. Przegląd elektrotechniczny. 2019. vol. 95, no 7. S. 38–41.

[10] Tao Zhan, Kun Yin, Jianghao Xiong, Ziqian He, Shin-Tson Wu, Augmented Reality and Virtual Reality Displays: Perspectives and Challenges, iScience, Volume 23, Issue 8, 2020, 101397, ISSN 2589-0042, https://doi.org/10.1016/j.isci.2020.101397.

[11] Margarido, Solange & Cardoso, Jorge. (2019). Web-based Virtual Reality with A-Frame. 1-2. 10.23919/CISTI.2019.8760795

[12] X. Huang, "Research and Application of Node.js Core Technology," 2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), Sanya, China, 2020, pp. 1-4, DOI: 10.1109/ICHCI51889.2020.00008.

An Investigation into the Efficiency of Specific Databases for Tracking Purposes in Scope of IT Startup

Volodymyr A. Franiv
Optoelectronics and Information Technologies Department
Ivan Franko National
University of Lviv
Lviv, Ukraine
volodymyr.franiv@lnu.edu.ua

Oleksandr R. Biletskyi
Ivan Franko National
University of Lviv
Lviv, Ukraine
oleksandr.biletskyi@lnu.edu.ua

Sofiya V. Vasylyuk
Department of Technology of Biologically Active Substances,
Pharmacy and Biotechnology
Lviv Polytechnic National University
Lviv, Ukraine
sofiia.v.vasyliuk@lpnu.ua

Ihor A. Franiv
Faculty of Commodity, Management and Servicing
Lviv University of Trade and Economics
Lviv, Ukraine
ihor.franiv@lute.lviv.ua

***Abstract* — This science paper presents a comparative analysis of the performance of a location tracking application developed by our team using the ASP.NET framework. The study focuses on evaluating the application's performance based on different databases utilized for storing and retrieving location coordinates from multiple resources. The databases examined in this research include MySQL, MSSQL, PostgreSQL, Redis and MongoDB, representing a mix of relational and non-relational databases.**

The primary objective of this investigation is to offer guidance to startups faced with the decision of selecting an optimal database solution for tracking tasks within their projects. Given the resource limitations typically encountered by IT startups, maximizing database performance in terms of storing and managing data is of paramount importance.

Throughout the study, various performance metrics were measured and compared for each database option. By thoroughly assessing the performance of the application under different database, we aim to provide valuable insights into which database type would be most suitable for specific tracking tasks in startup projects.

The results of this analysis shed light on the strengths and weaknesses of each database, highlighting the potential trade-offs that startups might encounter when choosing a particular database for their tracking applications. Ultimately, the findings presented in this article aim to empower startups in making informed decisions that align with their specific project requirements, ensuring they can leverage the maximum performance from their selected database while efficiently managing stored data.

Keywords* — *databases, throughput, tracking, MsSQL, PostgreSQL, MySQL, Redis, MongoDB

I. Introduction

In today's rapidly evolving world, the task of location tracking has emerged as a critical and ubiquitous requirement across numerous domains, including logistics, agriculture, and military operations [1]. The ability to efficiently receive and store location coordinates from multiple resources has become paramount in optimizing processes, enhancing security, and making well-informed decisions. As a result, the demand for reliable and high-performance tracking solutions has grown exponentially.

In the realm of Information Technology, startups are continually seeking innovative and robust technologies to build their applications [2-3]. Open-source solutions [4-7], with their flexibility and cost-effectiveness, have become increasingly popular among these burgeoning ventures. However, with a plethora of options available, selecting the most suitable technology stack, particularly for data storage, can be a daunting challenge.

Among the various data storage solutions, databases [8-11] stand out as one of the most vividly employed technologies for managing tracking data. Whether it involves tracking the movement of goods in logistics, monitoring crop growth in agriculture, or ensuring the security of military assets, databases play a crucial role in efficiently capturing, organizing, and retrieving location information.

This science paper delves into a comparative analysis of the performance of a location tracking application that our team developed using the ASP.NET framework [4]. The study focuses on evaluating the application's performance based on different databases [10-13] employed to store and retrieve location coordinates. Our goal is to provide startups with valuable insights into the optimal database solution for their tracking tasks, considering the relevance of tracking in diverse industries and the compelling need for open-source, robust technologies in the IT startup landscape. Similar research was done by [14] but for application developed using java and doesn't cover Redis database.

Throughout this article, we will explore and assess the efficiency of various databases, ranging from traditional relational databases like MySQL and MSSQL, PostgreSQL to modern non-relational databases like MongoDB and Redis [9-11]. By investigating the performance of application, we seek to offer practical advice to startups, enabling them to make informed decisions about the most appropriate database solution to maximize performance and effectively manage their valuable tracking data.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

In the subsequent sections, we will delve into the research methodology, present the database comparison results, and discuss the implications of our findings on the choice of the most suitable database for tracking tasks in startup projects that deices to build their application using .NET. By addressing this critical aspect of application development, we aim to contribute to the growth and success of startups in diverse industries, facilitating their ability to harness the full potential of location tracking technology.

II. Experiments and results

A. Tracking application, and performance measurements

In the pursuit of a robust and efficient tracking solution for various resources, our team embarked on the development of a cutting-edge location tracking application utilizing the powerful ASP.NET framework [5-6]. The application was designed to handle HTTP requests, with a primary focus on efficiently receiving and storing location coordinates for resources in real-time. With seamless integration of the HTTP PUT method, the application allowed for the addition of new coordinates as they were received, forming a continuous sequence of resource locations. The ingenious simplicity of this approach ensured that no data points were missed, providing a comprehensive and accurate track of each resource's movement.

One of the core functionalities of our application lies in its ability to cater to distinct HTTP GET requests. First, we implemented a request that fetched the latest coordinate of a resource from different databases (last position). Despite the absence of explicit indexes, the .NET framework's inherent efficiency ensured swift retrieval of the most recent data, facilitating real-time monitoring of resource locations. Additionally, we incorporated another GET request tailored to retrieve the complete track of a resource. Leveraging the ASP.NET framework's exceptional database handling capabilities, this feature enabled users to obtain a comprehensive history of a resource's movement with ease.

Choice of ASP.NET for IT Startup: In the dynamic landscape of IT startups, the decision to build our tracking application on the .NET framework was not taken lightly[5]. We recognized that the success of a startup hinges on several key factors, including rapid development, cost-effectiveness, and scalability. The ASP.NET framework emerged as an ideal solution that perfectly aligned with these startup-centric prerequisites. Its extensive libraries, pre-built components, and developer-friendly environment significantly accelerated our application development process [6].

Furthermore, the inherent compatibility of .NET with various operating systems and web servers ensured a seamless deployment experience [4]. With ASP.NET's open-source nature and active community support, we could harness its flexibility and optimize our tracking application to meet the specific requirements of diverse industries. This pivotal choice of the ASP.NET framework has laid a solid foundation for posible startup, empowering us to focus on delivering an exceptional tracking solution while keeping overhead costs at a minimum.

In summary, our location tracking application developed on the ASP.NET framework [5] stands as a testament to its prowess in the context of IT startups. Its unmatched ability to handle HTTP requests, store location data efficiently, and provide real-time tracking capabilities has positioned our solution at the forefront of the tracking industry.

To ensure the robustness and efficiency of our location tracking application, we leveraged the powerful capabilities of Apache Jmeter [7], a renowned open-source performance testing tool. With its versatility and user-friendly interface, Apache JMeter became an indispensable asset in our development process.

1. Request Generation with Apache JMeter

Apache Jmeter [7] played a pivotal role in generating a wide array of HTTP PUT and GET requests to thoroughly evaluate the capabilities of our tracking application. Leveraging JMeter's intuitive interface, we crafted custom scenarios and simulated real-world usage patterns to comprehensively test the application's performance under varying loads.

For HTTP PUT requests, we simulated the influx of new location coordinates from multiple resources by configuring JMeter to generate a stream of data points. This allowed us to observe how our application efficiently processed incoming data in real-time and maintained a seamless sequence of resource locations. By executing these extensive tests, we ensured that no data points were missed or mishandled, guaranteeing the accuracy and reliability of the tracking system.

Similarly, for HTTP GET requests, we utilized Apache JMeter to simulate diverse scenarios. The first type of GET request focused on retrieving the latest coordinate of a resource from different databases. By subjecting our application to varying levels of data load, we gauged its responsiveness in swiftly retrieving the most recent location data. This aspect was particularly crucial in enabling real-time monitoring of resources, ensuring that stakeholders could access up-to-date information promptly.

The second type of GET request involved retrieving the complete track of a resource. Through JMeter's capabilities, we emulated scenarios where users accessed historical data, and the application had to fetch and present extensive tracks. This exercise provided essential insights into the application's ability to handle large data sets efficiently, facilitating seamless access to resource movement history.

All tests were performed on local machine with 8 GB RAM, Intel Core i5 processor, operation system: Ubuntu Linux 18.01.

2. Measuring Application Performance with Apache JMeter

In addition to request generation, Apache JMeter served as a comprehensive performance measurement tool. We meticulously designed performance test plans, replicating diverse usage scenarios and load conditions. Through JMeter's extensive reporting and analysis features, we gathered throughput as a main metrics for our experiments.

By simulating various levels of concurrent users and resource tracking activities, we stress-tested our application to assess its scalability and responsiveness under heavy loads.

The insights gleaned from Apache JMeter's performance testing enabled us to fine-tune our application and enhance its overall efficiency, ensuring that it could handle a multitude of tracking requests without compromising on speed and accuracy.

Experimental Approach: Evaluating Application Bandwidth under Gradual Data Loading

In this study, we aimed to comprehensively assess the bandwidth of our location tracking application in response to varying data loads. The experiment focused on measuring the number of successfully processed HTTP GET requests within a one-second interval, representing the application's throughput under different data influx scenarios. The key novelty of our approach lay in the deliberate omission of special indexes in the database, allowing us to gauge the inherent efficiency of the application in managing tracking data without additional indexing support.

Gradual Data Loading: Simulating Real-World Scenarios

To replicate real-world usage patterns, we systematically increased the volume of data in the database using a series of HTTP PUT requests. With each PUT request, new location coordinates were seamlessly added to the system, mimicking the continuous inflow of data observed in real-time tracking scenarios. This incremental approach enabled us to examine the application's bandwidth across varying levels of data load, shedding light on its ability to handle progressively increasing data volumes.

An Emphasis on Database Efficiency: Absence of Special Indexes

A critical aspect of our experimental design was the deliberate choice not to create any special indexes in the database. This approach would minimize costs spend on manage database which is crucial for IT startup. Also we sought to isolate and evaluate the raw performance of the application without relying on additional indexing mechanisms. This allowed us to focus solely on the database handling capabilities inherent to the application's architecture. The absence of special indexes ensured that the results directly reflected the application's intrinsic efficiency in processing HTTP GET requests.

Consistency and Reliability: A Standardized Test Environment

To ensure the reliability and validity of our findings, all experiments were conducted on the same personal computer with consistent resource allocation. This standardized test environment minimized potential variations and confounding factors that could influence the results. By performing the experiment under controlled conditions, we aimed to provide a robust basis for qualitative analysis, enabling a nuanced understanding of the application's bandwidth in response to increasing data loads.

B. Relation databases

We conducted experiments to evaluate the performance of our application with three commonly utilized relational databases: MySQL 8.0.11, PostgresSQL 13.2, and MSSQL 15.0 [14-20]. Regarding the database schemas for relational databases, we examined two possible options. The first schema involved a single table that stored data for all resources (table 1), while the second schema utilized multiple tables, each dedicated to a specific resource for storing related data exclusively.

TABLE I. TABLE SNIPPET WITH DATA STORED IN RELATION DATABASE

Resource identifier	Longitude	Latitude	Timestamp
1	-77.0365	38.8951	2023-05-04 12:34:28
2	-77.0366	38.8952	2023-05-04 12:34:29
3	-77.0367	38.8953	2023-05-04 12:34:30
4	-77.0368	38.8954	2023-05-04 12:34:31
….			
100000	-78.958	39.9227	2023-05-04 21:24:13

The performance measurement results are depicted in figures 1 to 3

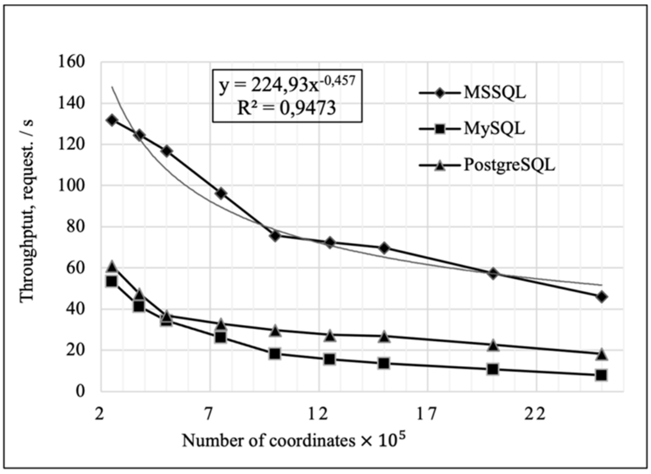


Fig. 1. Dependence of throughput (get track for resource) on number of coordinates, stored in different relation databases. Database with single table.

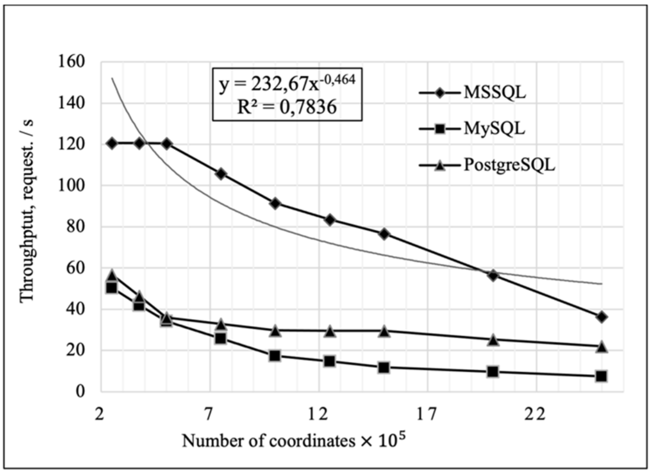


Fig. 2. Dependence of throughput (get last resource coordinate) on number of coordinates, stored in different relation databases. Database with single table.

It is evident that across all tested cases, for both databases and database schemas, there is a consistent pattern: the throughput decreases as the number of coordinates stored in the corresponding database increases [19]. Notably, from figure 1, it is apparent that the most optimal performance is achieved when using our developed application in conjunction with MSSQL. In the same graph, it is evident that MySQL and PostgresSQL exhibit similar throughput values.

From figure 2, we can observe a similar curve, which can be best explained by a power function. It is worth highlighting that the throughput values for the cases "get last resource coordinate" and "get resource track" are very close. Moreover, the throughput for MySQL is approximately three points higher than that of PostgresSQL.

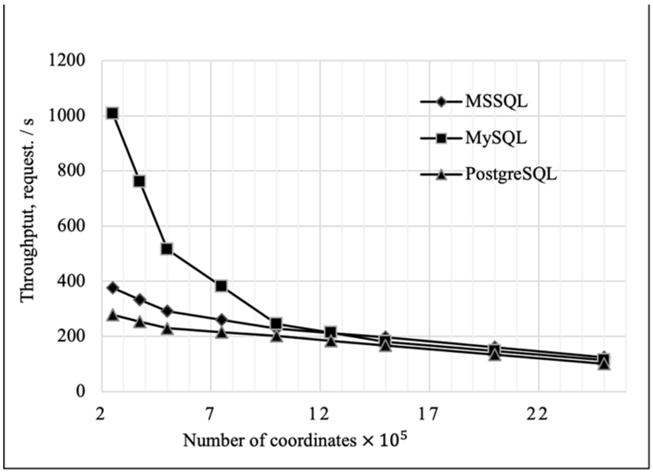


Fig. 3. Dependence of throughput (get resource track) on number of coordinates, stored in different relation databases. Database with multiple tables.

Figure 3 demonstrates that adopting a database schema with multiple tables results in significantly higher throughput values, nearly reaching 1000 requests per second, compared to the best case of 130 requests per second for the MySQL database. Additionally, the throughput values for MySQL and PostgresSQL are also approximately 100 times higher. The observed behavior follows a similar trend and can be explained by a power function. These results align with previous research in the field, where studies have shown that adopting multiple tables for database schema can lead to improved performance and efficiency, particularly when dealing with large datasets and complex queries [12-16]. Additionally, other studies have also highlighted the significance of considering throughput as a performance metric when comparing relational databases in real-world applications[14].

C. Non-relational databases

In this experiment, we investigated widely used non-relational databases, specifically MongoDB 6.0 and Redis DB 5.0 [16-20]. In the case of MongoDB, the database comprises documents with the following fields: id, longitude, latitude, and timestamp. On the other hand, for Redis DB, we employed sorted sets as a data structure to store resource coordinates. These sorted sets were organized based on the date of each coordinate, resulting in a collection of sorted sets, each dedicated to a particular resource. Sets were distinguished by resource ID.

As depicted in Figure 4, the behavior exhibited by these non-relational databases closely resembles that of relational databases. The observed curve can be better understood through a power function explanation. Notably, the throughput values for MongoDB are in proximity to those of relational databases. However, a significant distinction is apparent in the case of Redis. Here, we observe substantially higher throughput values, approximately 15 times greater. This discrepancy could likely be attributed to the fact that Redis utilizes RAM for data storage. A distinctive behavior, illustrated in Figure 5 for Redis DB, presents a more intricate aspect that demands further investigation for proper understanding.

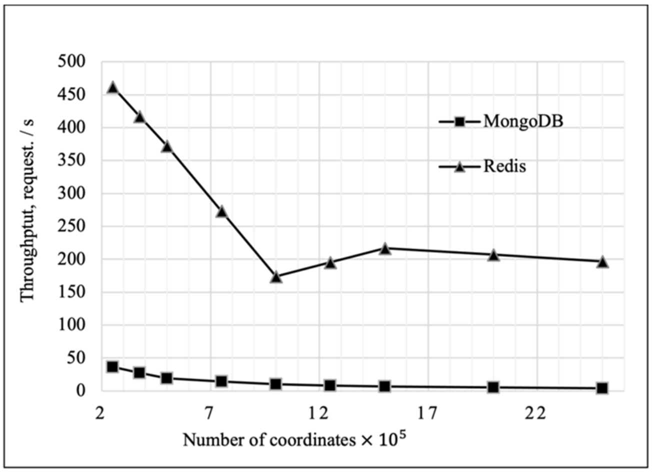


Fig. 4. Dependence of throughput (get resource track) on number of coordinates, stored in different non-relation databases.

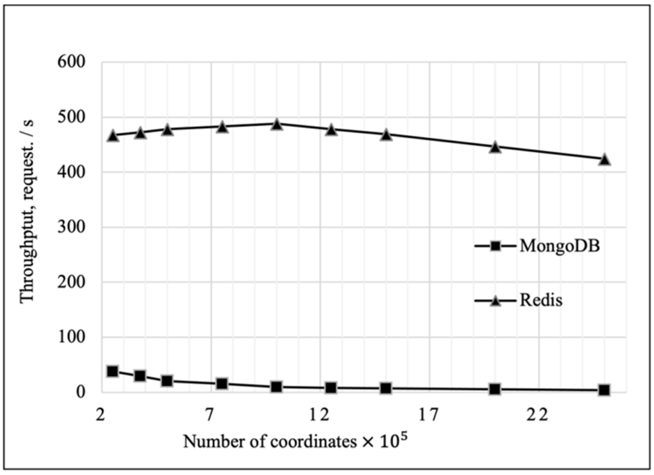


Fig. 5. Dependence of throughput (get last coordinate) on number of coordinates, stored in different non-relation databases.

CONCLUSIONS

In conclusion, this scientific inquiry delved into a comprehensive comparative analysis of the efficiency of specific databases for tracking purposes within the context of a location tracking application developed using the .NET framework. By meticulously examining both relational and non-relational database options, our study sought to address a critical concern for startups – the optimal selection of a database solution to maximize performance and effectively manage tracking data.

Our investigation illuminated crucial insights into the performance dynamics of various databases. The comparative analysis of relational databases, including MySQL, PostgreSQL, and MSSQL, demonstrated nuanced variations in throughput, response times, and scalability. Notably, the .NET framework showcased its prowess in swiftly processing HTTP requests and storing location data efficiently, solidifying its status as a robust choice for startups seeking rapid development and scalability.

The assessment of non-relational databases brought forth valuable findings as well. MongoDB's throughput proximity to relational databases coupled with Redis DB's remarkable performance highlighted the significance of non-relational solutions, particularly in scenarios demanding substantial data storage and retrieval.

Based on our research among relation databases we observe highest performance for MSSQL and Redis DB in case of non-relation databases.

As startups strive to navigate the complex landscape of technology solutions, this study's outcomes provide a clear guide for decision-making. By coupling the strengths of the .NET framework with appropriate database choices, startups can enhance their tracking applications' speed, accuracy, and overall efficiency.

REFERENCES

[1] F. Ahmed, M. Phillips, S. Phillips, K.-Y. Kim, "Comparative Study of Seamless Asset Location and Tracking Technologies," Procedia Manufacturing, vol. 51, 2020, pp. 1138–1145.

[2] M. Akkaya, Startup Valuation, IGI Global, 2019, pp. 137–156,

[3] A. Mathur, H. Agarwal, "Study of Challenges Faced by Startup Industries," A referred & peer- reviewed quarterly research journal. Vol 48, 2023, pp 58-67.

[4] D. Esposito, Building Web Solutions with ASP.NET and ADO.NET. Redmond: Microsoft Press, 2002.

[5] K. Padaliya, C# Programming with .Net Framework, 2019.

[6] M. Smorgun, ".NET Tools for Software Development: Tool Selection, Key Benefits of .NET Web Applications," Asian Journal of Research in Computer Science, 2023, 15(2):43-56.

[7] S. Matam, J. Jain, Pro Apache JMeter: web application performance testing. Apress. USA. 2017.

[8] C. H. Lee, Y. L. Zheng, "SQL-to-NoSQL Schema Denormalization and Migration: A Study on Content Management Systems," In: Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC) , 2015, pp. 2022–2026.

[9] M. Abu Kausar, M. Nasar, A. Soosaimanickam, "A Study of Performance and Comparison of NoSQL Databases: MongoDB, Cassandra, and Redis Using YCSB." Indian J. Sci. Technol 15, 2022, pp. 1532-1540.

[10] J. Lourenco, B. Cabral, P. Carreiro, M. Vieira, J. Bernardino, "Choosing the right NoSQL database for the job: a quality attribute evaluation," In: Journal of Big Data 2, 2015, pp. 1–26.

[11] L. Vokorokos, M. Uchnar, L. Lescisin, "Performance optimization of applications based on non-relational databases," In: International Conference on Emerging eLearning Technologies and Applications (ICETA), 2016, pp. 371–376.

[12] N. Jatana, S. Puri, M. Ahuja, I. Kathuria, D. Gosain, "A survey and comparison of relational and non-relational database," International Journal of Engineering Research & Technology, 1(6), 2012.

[13] J. Han, E. Haihong, G. Le, J. Du, "Survey on NoSQL database," In 6th International Conference on Pervasive Computing and Applications (ICPCA), 2011, pp. 363-366.

[14] K. Fraczek, M. Plechawska-Wojcik "Comparative analysis of relational and non-relational databases in the context of performance in web applications," International Conference: Beyond Databases, Architectures and Structures, 2017.

[15] S. Gupta, G. Narsimha, "Efficient Query Analysis and Performance Evaluation of the Nosql Data Store for BigData," Proceedings of the First International Conference on Computational Intelligence and Informatics. Springer Singapore, 2017, pp. 549–558.

[16] K. Chodorow, M. Dirolf, MongoDB: The Definitive Guide (1st ed.), O'Reilly Media, 2010.

[17] D. Sullivan, NoSQL for Mere Mortals. Addison-Wesley, 2015.

[18] E. Brewer, CAP twelve years later: How the rules have changed. In: Computer 45,2 2012, pp. 23–29.

[19] X. Li, Z. Ma, H. Chen, "QODM: A query-oriented data modeling approach for NoSQL databases:," Advanced Research and Technology in Industry Applications (WARTIA), 2014, pp. 338–345.

[20] Y. Li, S. Manoharan, "A performance comparison of SQL and NoSQL databases," Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM) 2013, pp. 15-19.

[21] C. Truica, O. Radulescu, F. Boicea, A. Bucur, "Performance evaluation for CRUD operations in asynchronously replicated document oriented database," Proceedings of 20th International Conference on Control Systems and Computer Science, 2015, pp. 191–196.

Comparative Study of ABC and GWO Implementations on Raspberry Pi 3

Oleh Sinkevych
Electronics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
oleh.sinkevych@lnu.edu.ua

Yaroslav Boyko
Electronics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
yaroslav.boyko@lnu.edu.ua

Bohdan Sokolovskii
Electronics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
bohdan.sokolovskyy@lnu.edu.ua

Igor Olenych
Electronics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
igor.olenych@lnu.edu.ua

Liubomyr Monastyrskii
Electronics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
lyubomyr.monastyrskyy@lnu.edu.ua

Mykhailo Pavlyk
Electronics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
mykhailo.pavlyk@lnu.edu.ua

***Abstract* — The paper provides a comparative analysis of two popular swarm meta-heuristic algorithms deployed on the Raspberry Pi 3 microcomputer. Both Artificial Bee Colony (ABC) and Grey Wolf Optimizer (GWO) were implemented from scratch in pure C in order to study their performance on relatively low-power device. Due to the limited computing power of Raspberry Pi 3, it is important to evaluate its productiveness in the execution of ABC and GWO to figure out the corresponding computational capabilities. For this task, a bunch of complicated benchmark functions have been used. The numerical experiments have shown both the advantages and disadvantages of these methods running on RPi3, as well as the peculiarities of the configurations that affect the time of complete execution of the tasks. It has been quantitatively discovered, that the increase of problems complexity leads to the significant increase of execution time. The obtained results signify the necessity of boosting the performance by using the concurrent computations which is simplified by the availability of four physical CPU cores and suitability of the considered algorithms for parallelization.**

Keywords — Artificial bee colony, grey wolf optimizer, Raspberry Pi 3, numerical optimization, swarm intelligence

I. Introduction

Swarm algorithms try to mimic the organizational and search strategy of collective living beings. This makes it possible to effectively solve many applied problems [1]. Their taxonomy includes, in particular, artificial ant and bee colonies, wolf and bird packs, lion prides, blue whales, and bacteria foraging [2]. Their relatively good efficiency is explained by the adaptation of the evolutionary experience of various strategies in harsh environmental conditions. The research and implementation of these methods allow not only to solve the problems of finding the optimal path or optimizing the function but also to build artificial swarm systems like unmanned robotic complexes.

Artificial bee colony (ABC) demonstrates the ability to effectively solve complex engineering problems as one of the best representatives of the mentioned family of algorithms. Developed by D. Karaboga [3], it is still one of the widely used methods in numerical optimization. Various studies have been done in order to enhance its performance. For instance, in [4-6] authors proposed different techniques to boost ABC capabilities in solving constrained and unconstrained optimization problems.

Another well-known representative of swarm algorithms is grey wolf optimizer (GWO), whose idea is based on the attacking tactics of a pack of wolves. It is one of the most cited methods of swarm meta-heuristics [7]. Despite the fact that this method is primarily aimed at solving typical optimization problems, it contains the idea of synchronizing wolves-agents based on the hierarchy of the pack. This opens up the possibility of its implementation, for instance, in drone swarm organization approaches. Unfortunately, a generic GWO algorithm suffers from considerable defect [8], but nevertheless it is still suitable for improvements [9, 10].

We selected these two algorithms as outstanding and popular instances of swarm intelligence, since the interaction scheme of search agents can be reproduced in the context of the interaction of hardware systems endowed with artificial intelligence. In addition, both algorithms do not require many hyper-parameters, which simplifies their configuration.

To determine the starting point of such research, it is necessary to analyze both the capabilities of the hardware as a component of the swarm system, and the performance of these algorithms in the basic optimization scenarios. The analysis of scientific publications showed that the deployment and research of swarm algorithms on embedded devices has received extremely little attention. The reasons for this are a relatively large number of calculations of the objective function, which represents the problem and complexity of scaling the algorithm. Accordingly, the paper investigates the work of ABC and GWO (including their modifications) algorithms on the Raspberry Pi 3 in the sense of solving unconstrained optimization problems.

II. ABC and GWO Concepts

In this section, we provide the core ideas of ABC and GWO algorithms as well as the corresponding mathematical

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

statements. To start with, let's consider the unconstrained optimization problem

$$\mathbf{x}^* = \arg\min F(\mathbf{x}), \mathbf{x} \in \Re^n, \qquad (1)$$

where $\mathbf{x}$ is the n-dimensional vector and F is the objective function which represents the problem, $\mathbf{x}^*$ is the global minimum. Such a problem can be generally related to finding some optimal path, scheduling workloads and even training or fine-tuning the neural network. In case of constraints for the function or vector values, they can be easily added to (1).

Both ABC and GWO are aimed at solving (1) by applying different search methods. Primarily, this process starts with generation of N initial random solutions via

$$x_i^j = l^j + \varphi\left(u^j - l^j\right), \qquad (2)$$

where x_i^j is the j-th component of vector $\mathbf{x}_i$, $i \in [0,...,N-1]$, j is the index of dimension, $\varphi = rand(0,1)$ is the fandom value in $(0,1)$, l^j and u^j are lower and upper bounds in j-th dimension. We should remark that (2) can be modified, e.g., using Levy distribution [11].

After this, specific techniques defined within ABC and GWO framework are applied to evolve $\mathbf{x}_i$ towards the best solution of (1).

A. Artificial Bee Colony

The Artificial Bee Colony algorithm is inspired by bee's foraging mechanics, when a bunch of agents are seeking the best food source while keep communicating with each other. In order to solve (1), after executing (2), ABC performs a set of steps. These steps can be represented as follows:

1. *Setup*. Each solution $\mathbf{x}_i$ is mapped to fitness function value $fit(\mathbf{x}_i)$ and to the trial counter C_i, which signifies that the number of times $\mathbf{x}_i$ has not been improved. The fitness value $fit(\mathbf{x}_i) = 1 + |F(\mathbf{x}_i)|$, if $F(\mathbf{x}_i) < 0$ or $fit(\mathbf{x}_i) = 1/(1 + F(\mathbf{x}_i))$, if $F(\mathbf{x}_i) \ge 0$. The number of abstract search agents (bees) equals number of solutions N.
2. *Employed Bee Phase*. During this phase, bees mutate solution $\mathbf{x}_i$ as follows:

$$x_i^j = x_i^j + \phi\left(x_i^j - x_k^j\right), k \ne j, \qquad (3)$$

where x_k^j is the component of randomly chosen solution k, $\phi = rand(-1,1)$. If $fit\left(\mathbf{x}_i\left(..., x_k^j, ...\right)\right)$ does not get better, than this solution becomes abandoned and $C_i + = 1$.
3. *Onlooker Bee Phase*. After the employed bee phase, calculation of probabilities is done via $P_i = 0.1 + 0.9\left(fit(\mathbf{x}_i)/\text{maxFit}\right)$, where maxFit is the highest fitness value among the all. Then every onlooker bee $\in [0,...,N-1]$ improves solutions based on probabilities: if $rand(0,1) < P_i$, then employed bees step is performed. The latter executes until all onlookers do improvement.
4. *Scout Bee Phase*. Scout simply checks the trials C_i and if reaches, for instance, $(N \cdot n)/2$, then reinitializes solution $\mathbf{x}_i$ using (2).

Steps 2-4 are repeated predefined *maxIteration* times. If during the search process some solution's component runs over range l^j or u^j, then it have to be returned via assigning l^j or u^j value.

An interesting point here is that these steps can be implemented in parallel which increases execution time [12]. There is practically no interaction between agents, which allows more efficient use of multithreading. When designing the swarm system, this does not determine the appealing mechanism of communication. However, it can be a basis for creating the effective search algorithm.

B. Grey Wolf Optimizer

Unlike ABC, in GWO the communication between agents is more noticeable. The steps of GWO consist of

1. *Setup*. The population of wolves (*w.r.t.* solutions) is generated via (2) and fitness values $F_i = F(\mathbf{x}_i)$ are calculated; parameter $a = 2$ which is linearly decreasing down to 0 during the main loop and maximum number of iterations are set in advance.
2. *Leaders updates*. Among the all $\mathbf{x}_i$ three best solutions $\mathbf{X}_\alpha$, $\mathbf{X}_\beta$ and $\mathbf{X}_\delta$ are obtained based on their fitness values. These solutions represent α, β and δ wolves in the pack which lead other (ω) members toward the prey.
3. *Hunting process*. For each wolf i and for each dimension j three distances to prey are calculated

$$D_k = \left|C_k x_k^j - x_i^j\right|, k \in \{\alpha, \beta, \delta\}, \qquad (4)$$

where $C_k = 2 \cdot rand(0,1)$, $x_k^j \in \mathbf{X}_k$, and new leader's positions X_k are updated as follows:

$$X_k = x_k^j - A_k \cdot D_k, \qquad (5)$$

where $A_k = 2a \cdot rand(0,1) - a$. This step moves each leader closer and closer to prey (optimal solution). The parameter A_k stands for exploitation and exploration purpose, i.e., if $|A_k| < 1$, then wolves move (attack) towards the prey, otherwise wolves diverge from the prey to explore new solution.
4. *Solution update.* Based on the leader positions X_k, the new solution x_i^j is calculated as averaged value of

X_k : $x_i^j = \left(X_\alpha + X_\beta + X_\delta\right)/3$, i.e., three leaders define new solution.

5. *Verification*. If the new solution does not get better than previous, values of the latter are kept instead. Also, the checking of bounds for each produced solution's component should be done.

As in ABC, here steps (2-5) are repeated $maxIteration$ times.

Due to the significant limitation of GWO [8], the generic algorithm fails to effectively deal with functions whose global minimum is reached at a point far from zero. Hence, a lot of modifications have been proposed to overcome this issue. For this research we selected relatively simple modification MGWO [13], where instead of selection three best leaders, the random ones are chosen every iteration among the all solutions (wolves).

A concept behind GWO algorithm forms a firm basis for the development of communicational mechanism for swarm (multi-agent) system.

III. Numerical Experiments

Raspberry Pi 3 B+ microcomputer is considered as cheap and high-quality hardware to be a "brain" unit for autonomous devices. For instance, 1.4GHz 64-bit quad-core processor, dual-band wireless LAN, Bluetooth 4.2/BLE are good indicators for the deployment and research of the software base of swarm intelligence system. To measure RPi3 hardware capabilities in performing ABC and GWO/MGWO steps, we selected six benchmark functions (Table I) to minimize, where λ is a shift parameter. For all the functions $F\left(\mathbf{x}^*\right) = 0$.

TABLE I. Benchmark functions

Name	Function
Sphere	$\sum_{i=1}^{D}(x_i-\lambda)^2, \lambda=50, x_i \in [-100;100]$
Rastrigin	$10D+\sum_{i=1}^{D}\left[(x_i-\lambda)^2-10\cos\left(2\pi(x_i-\lambda)\right)\right], \lambda=2,$ $x_i \in [-5.12;5.12]$
Schwefel	$418.98d-\sum_{i=1}^{D}x_i\sin\left(\sqrt{\lvert x_i\rvert}\right), x_i \in [-500;500]$
Griewank	$\sum_{i=1}^{D}\frac{(x_i-\lambda)^2}{4000}-\prod_{i=1}^{D}\cos\left(\frac{x_i-\lambda}{\sqrt{i}}\right)+1, \lambda=50, x_i \in [-100;100]$
Rosenbrock	$\sum_{i=1}^{D-1}\left[100\left(x_{i+1}-x_i^2\right)+(x_i-1)^2\right], x_i \in [-10;10]$
Ackley	$20+\exp-20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D}(x_i-\lambda)^2}\right)-$ $\exp\left(\frac{1}{D}\sqrt{\sum_{i=1}^{D}\cos\left(2\pi(x_i-\lambda)\right)}\right), \lambda=10, x_i \in [-32;32]$

The dimensions D for each function are $D_i = \{50, 100, 200\}$. The number of search agents for all algorithms (number of solutions $\mathbf{x}_i$) is set to 50; $maxIteration = 5000$. All the algorithms were implemented in pure C programming language.

A. ABC results

We have conducted 5 sequential numerical experiments for each dimension from D_i and measured execution time t_{ex} with the corresponding averaged function values $F\left(\mathbf{x}^*\right)$ for each experiment and the number of iteration related to found minimum with pre-defined accuracy $\varepsilon = 0.01$. The results are shown in Tables 2-5.

TABLE II. ABC results for D=50

Name	t_{ex}	$F\left(\mathbf{x}^*\right)$	Iterations
Sphere	10	1.00E-03	476
Rastrigin	37	1.00E-03	1707
Schwefel	41	1.00E-03	2909
Griewank	26	4.00E-02	704
Rosenbrock	122	1.00E-03	4574
Ackley	28	1.00E-03	1049

TABLE III. ABC results D=100

Name	t_{ex}	$F\left(\mathbf{x}^*\right)$	Iterations
Sphere	42	1.00E-03	973
Rastrigin	174	1.00E-03	3895
Schwefel	126	3.50E+02	5000
Griewank	102	2.00E-01	1314
Rosenbrock	359	1.00E-03	5000
Ackley	117	1.00E-03	2132

TABLE IV. ABC results D=200

Name	t_{ex}	$F\left(\mathbf{x}^*\right)$	Iterations
Sphere	184	1.00E-03	2061
Rastrigin	468	1.60E+01	5000
Schwefel	246	4.74E+03	5000
Griewank	364	1.18E+02	2171
Rosenbrock	952	1.00E-03	5000
Ackley	533	1.00E-03	4539

As it is seen from Table II, the ABC algorithm allows to solve all seven problems in the average for 44 seconds and 1903 iterations, where single iteration lasts 0.02 second. With the increase of dimensions up to 100 (Table III), the results for complicated Schwefel and Rosenbrock functions get worse, nevertheless are still close to real global minima. Here the average execution time equals 153 seconds and average number of iterations equals 3052, while one iteration lasts 0.05 second. Although previous results are quietly acceptable, the increase in dimensionality (Table IV) caused the significant deterioration within the given $maxIteration$ and number of agents. The only three problem were successfully solved; the average execution time equals 458 seconds, while average number of iterations grows up to 3952. Also, one iteration exceeds 0.12 second. When the dimensionality doubles, the average running time increases by about a factor of 3. The increase of $maxIteration$ and number of agents may lead to better results, but the overall execution time will be higher.

B. GWO and MGWO results

We have repeated the numerical experiments on seven benchmark functions using two grey wolf optimizers: the first one is generic original implementation and the second one is modified version [13]. The results of GWO and MGWO are presented in tables 5-7 and 8-10 correspondingly.

TABLE V. GWO RESULTS FOR D=50

Name	t_{ex}	$F(\mathbf{x}^*)$	Iterations
Sphere	32	2.20E+04	5000
Rastrigin	38	1.21E+02	5000
Schwefel	33	1.12E+03	5000
Griewank	44	4.60E+01	5000
Rosenbrock	39	6.00E-01	5000
Ackley	38	4.30E+00	5000

TABLE VI. GWO RESULTS FOR D=100

Name	t_{ex}	$F(\mathbf{x}^*)$	Iterations
Sphere	67	2.50E+04	5000
Rastrigin	76	3.14E+02	5000
Schwefel	65	2.50E+04	5000
Griewank	87	9.60E+01	5000
Rosenbrock	77	6.00E+00	5000
Ackley	76	1.30E+01	5000

TABLE VII. GWO RESULTS FOR D=200

Name	t_{ex}	$F(\mathbf{x}^*)$	Iterations
Sphere	144	1.20E+05	5000
Rastrigin	151	7.40E+02	5000
Schwefel	135	4.63E+04	5000
Griewank	190	1.98E+02	5000
Rosenbrock	158	2.90E+01	5000
Ackley	152	1.60E+01	5000

The obtained results clearly demonstrate that original GWO implementation has considerable issues with functions whose global minimum is reached in points far away from zero. This drawback was described in detail in [8]. Unfortunately, there is still a big question whether this effect can be reduced without complete rebuilding of GWO. Moving on to the results for 50 dimension problems only Griewank function was successfully minimized, while the rest functions were too challenging for GWO. The average execution time is approximately 37 seconds and the average number or iterations equals $maxIteration$. As the dimensionality increases to 100, the average accuracy of the results deteriorates by about two times. This can be explained both by "curse of dimensionality effect" and by getting stuck in local minima. The average time required to execute 5000 iterations rises up to 75 seconds. Increasing the dimensionality to 200 significantly worsens the accuracy; in this case the average running time exceeds 155 seconds.

In contrast to the original algorithm, MGWO version shows much better accuracy while neglects the core GWO idea, namely, to rely on the best three wolves in the pack. Regrettably, such a modification breaks the concept of hierarchical structure in a wolf pack. However, as the following results show, this approach adds diversity to the formation of new solutions through the participation of random wolves.

For the 50 dimension problem MGWO generally performs better than the original GWO. The average function values after 5000 iterations are lower while the average running time slightly increases to 45 seconds. For D equals 100 (average execution time is 88 seconds) and 200 (average execution time is 186 seconds) the dimension problems MGWO also produces better accuracy.

Both GWO and MGWO showed the worst results for complicated Schwefel problem.

TABLE VIII. MGWO RESULTS FOR D=50

Name	t_{ex}	$F(\mathbf{x}^*)$	Iterations
Sphere	42	1.00E-02	5000
Rastrigin	46	1.50E+01	5000
Schwefel	41	1.57E+04	5000
Griewank	50	4.50E+01	5000
Rosenbrock	45	1.00E-03	4980
Ackley	44	2.00E-02	5000

TABLE IX. MGWO RESULTS FOR D=100

Name	t_{ex}	$F(\mathbf{x}^*)$	Iterations
Sphere	81	1.00E-01	5000
Rastrigin	89	3.30E+01	5000
Schwefel	81	3.51E+04	5000
Griewank	95	1.03E+02	5000
Rosenbrock	92	1.00E-02	5000
Ackley	90	3.00E-02	5000

TABLE X. MGWO RESULTS FOR D=200

Name	t_{ex}	$F(\mathbf{x}^*)$	Iterations
Sphere	173	7.00E-01	5000
Rastrigin	193	1.07E+02	5000
Schwefel	159	7.36E+04	5000
Griewank	211	1.95E+02	5000
Rosenbrock	199	6.00E-02	5000
Ackley	180	6.00E-02	5000

IV. DISCUSSION AND CONCLUSIONS

The conducted comparative numerical experiments based on the described algorithms configurations on Raspberry Pi 3 platform have revealed the following derivations:

- Generic ABC implementation outperforms both GWO and MGWO in accuracy for three different dimension problems ($D_i = \{50,100,200\}$). However, the execution time of ABC is much higher (44 seconds ≥ (37 and 44) for 50 dimension problems, 153 seconds > (74 and 88) for 100 dimension problems, 458 seconds > (155 and 186) for 50 dimension problems).
- High execution time of ABC is caused by the onlooker phase of the algorithm; it can be modified or effectively parallelized.
- To speed up the ABC implementation on such a relatively slow hardware like Raspberry Pi 3 it is necessary to modify original ABC for multi-threading execution; using the properly implemented thread pool in C or configured OpenMP library can drastically decrease the execution time.

- Because pure C implementation is a bit low-code level and additional speed up techniques are complicated, it might be preferable to use C++ with the ability to integrate more high-performance tools like oneTBB [14].

- Generic GWO implementation, unfortunately, does not show very good accuracy and loses to ABC. The core problem here is defect found in [8]. Nevertheless, lower execution time and simplicity leave the possibility of its improvement in the direction of competition with ABC.

- Modified MGWO implementation behaves better than the original one, though it has not reached the ABC level of accuracy. Such a modification departs from the base idea of preserving hierarchy of leaders. Another possible improvement is the hybridization of different nature inspired approaches in order to overcome the defect common for all GWO based algorithms.

- Raspberry Pi 3 is able to execute both ABC and GWO steps even for high dimensional problems. Due to presence of 4 CPU cores these routines can be parallelized. Since one iteration of either ABC or GWO takes very little time (approx. 0.02 second), mechanics behind them can be easily implemented as a basis for communication between multi-agent hardware system.

References

[1] E. Cuevas, F. Fausto, and A. González, "Metaheuristics and Swarm Methods: A Discussion on Their Performance and Applications," in Intelligent Systems Reference Library. Cham: Springer International Publishing, 2019, pp. 43–67. Available: https://doi.org/10.1007/978-3-030-16339-6_2.

[2] A. E. Ezugwu et al., "Metaheuristics: A comprehensive overview and classification along with bibliometric analysis," Artif. Intell. Rev., vol. 54, no. 6, pp. 4237–4316, Mar. 2021. Available: https://doi.org/10.1007/s10462-020-09952-0.

[3] D. Karaboga, "Artificial bee colony algorithm," Scholarpedia, vol. 5, no. 3,p. 6915,2010.Available: https://doi.org/10.4249/scholarpedia.6915.

[4] Y. Shi, C.-M. Pun, H. Hu, and H. Gao, "An improved artificial bee colony and its application," Knowledge-Based Syst., vol. 107, pp. 14–31,Sep.2016. Available: https://doi.org/10.1016/j.knosys.2016.05.052.

[5] H. Gao, Y. Shi, C.-M. Pun, and S. Kwong, "An Improved Artificial Bee Colony Algorithm With its Application," IEEE Trans. Ind. Inform., vol. 15, no. 4, pp. 1853–1865, Apr. 2019. Available: https://doi.org/10.1109/tii.2018.2857198.

[6] S. Ghambari and A. Rahati, "An improved artificial bee colony algorithm and its application to reliability optimization problems," Appl. Soft Comput., vol. 62, pp. 736–767, Jan. 2018. Available: https://doi.org/10.1016/j.asoc.2017.10.040.

[7] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," Advances Eng. Softw., vol. 69, pp. 46–61, Mar. 2014. Available: https://doi.org/10.1016/j.advengsoft.2013.12.007.

[8] P. Niu, S. Niu, N. Liu, and L. Chang, "The defect of the Grey Wolf optimization algorithm and its verification method," Knowledge-Based Syst., vol. 171, pp. 37–43, May 2019. Available: https://doi.org/10.1016/j.knosys.2019.01.018.

[9] J.-S. Wang and S.-X. Li, "An improved grey wolf optimizer based on differential evolution and elimination mechanism," Scientific Rep., vol. 9, no. 1, May 2019. Available: https://doi.org/10.1038/s41598-019-43546-3.

[10] M. H. Nadimi-Shahraki, S. Taghian, and S. Mirjalili, "An improved grey wolf optimizer for solving engineering problems," Expert Syst. With Appl., vol. 166, p. 113917, Mar. 2021. Available: https://doi.org/10.1016/j.eswa.2020.113917.

[11] E. H. Houssein, M. R. Saad, F. A. Hashim, H. Shaban, and M. Hassaballah, "Lévy flight distribution: A new metaheuristic algorithm for solving engineering optimization problems," Eng. Appl. Artif. Intell., vol. 94, p. 103731, Sep. 2020. Available: https://doi.org/10.1016/j.engappai.2020.103731.

[12] D. Li, Y. Feng, J. Zhong, J. Zhou, L. Yin, and J. Zhou, "Parallel optimization based on artificial bee colony algorithm," in 2017 IEEE 2nd Int. Conf. Big Data Anal. (ICBDA), Beijing, China, Mar. 10–12, 2017.IEEE,2017.Available: https://doi.org/10.1109/icbda.2017.8078779.

[13] E. Akbari, A. Rahimnejad, and S. A. Gadsden, "A greedy non-hierarchical grey wolf optimizer for real-world optimization," Electron. Lett., vol. 57, no. 13, pp. 499–501, Apr. 2021. Available: https://doi.org/10.1049/ell2.12176.

[14] "GitHub - oneapi-src/oneTBB: OneAPI Threading Building Blocks (oneTBB)."GitHub. https://github.com/oneapi-src/oneTBB (accessed Aug. 18, 2023).

Implementation of Base Components of Neuro-like Cryptographic Data Protection Systems on FPGA

Ivan Tsmots
Department of Automated Control Systems
Lviv Polytechnic National University
Lviv, Ukraine
ivan.h.tsmots@lpnu.ua

Vasyl Rabyk
Department of Radiophysics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
vasyl.rabyk@lnu.edu.ua

Roman Tkachenko
Department of Publishing Information Technologies
Lviv Polytechnic National University
Lviv, Ukraine
roman.o.tkachenko@lpnu.ua

Yurii Opotyak
Department of Automated Control Systems
Lviv Polytechnic National University
Lviv, Ukraine
yurii.v.opotiak@lpnu.ua

Vasyl Teslyuk
Department of Automated Control Systems
Lviv Polytechnic National University
Lviv, Ukraine
vasyl.m.teslyuk@lpnu.ua

***Abstract* — The implementation of the base components of neuro-like cryptographic data protection systems using FPGA is considered. The structure of the data encryption module using polynomials based on a neuro-like network was developed. Module implemented with the help of VHDL in Quartus II development environment. The base components of different dimensions, which are part of the encryption module, have been implemented. The components allow quick implementation of encryption modules corresponding to different neural network architectures. Modeling of the module's operation and evaluation of the encryption time, hardware resources required for its implementation, and the implementation of the components included in the module were carried out. To prepare the parameters of a neural network of a given architecture, which are used in the implementation of the basic encryption operation, the Training_NS_MGT and Transfor_Data programs were implemented in the C language on the Raspberry Pi 4 microcomputer. Examples of the formation of matrices of weighting coefficients by these programs are given.**

Keywords* — *neural network, encryption/decryption of data using polynomials, basic encryption operation, FPGA EP3C16F484C6, development environment Quartus II ver. 13.1, simulation of module operation, time diagrams

I. Formulation of the Problem

Systems of cryptographic protection of data during information transmission are today a necessary component of any mobile platform. The need to provide remote control of such platforms requires the implementation of cryptographic protection in real time while taking into account limitations regarding dimensions, power consumption, and cost. One of the areas of modern cryptography is the use of artificial neural networks (ANNs). The paper [1] provides an overview of the use of different types of ANNs (Recurrent Neural Networks, general regression neural networks, chaotic neural networks, and multilayer neural networks) in cryptography, the advantages of which are a high degree of structural parallelism, their adaptive behavior, the ability to learn, generalize results, demonstrate high productivity. In particular, the works devoted to the synchronization of neural networks, cryptography based on chaotic neural networks, and multilayer neural networks were considered. When using the synchronization of neural networks, the secret key is not transmitted over an available channel but is generated with the help of neural are on the transmitting networks that and receiving sides. These neural networks are synchronized by a common external signal. The paper [2] presents a stream encryption system based on a pseudorandom number generator that uses ANN. The inverse error propagation algorithm was used to obtain the weighting factors in the ANN. After training, the ANN was hardware-implemented on the base of FPGA.

Most modern artificial neural networks (ANNs) are trained using iterative algorithms [1,2], which determine both their advantages and the main disadvantages and limitations during their applications. Disadvantages and limitations include the long training time of ANNs with a large number of weighting factors, and the unrepeatability of training results, which is associated with the random initialization of the weighting factors of the network. In [3, 4], auto-associative neural networks with a non-iterative principal component learning algorithm [5] were used for data encryption and decryption. The advantage of the method is the use of eigenvectors corresponding to the eigenvalues of the autocorrelation matrix. Neuro-like element is the base of Neural-like networks (NLN) and its functioning is reduced to scalar product calculation using the pre-calculated weighting coefficients. The algorithm of symmetric data encryption/decryption based on an auto-associative neural network is considered in [3]. The structure of the auto-associative network consists of input, hidden, and output layers. The learning algorithm consists in determining the matrix of weight coefficients of the autocorrelation matrix of the input data. Examples of the use of the algorithm for learning a neural network, encryption, and decryption of data based on it are given.

In the model of geometric transformations (MGT) [6] based neural networks non-iterative learning algorithm is used for a predetermined number of calculation steps, which ensures the repeatability of learning results. NLN based on MGT found its application in the implementation of cryptographic data protection methods. In [6], algorithms for learning NLN based on MGT with a teacher and without a teacher, which are simpler to implement and do not require the calculation of eigenvalues and eigenvectors, are discussed in detail.

To ensure small dimensions, power consumption, and cost of the neuro-like cryptographic data protection system, it is advisable to implement it on FPGA. In [7], a novel Izhikevich

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

neuron model is described which is a hardware realization and a low-cost implementation and may be used in a lightweight embedded cryptography system. As a proof-of-concept FPGA's realization is described. A design approach for a neural network proposed In [8] aimed to reduce the number of adders and multipliers in hardware. The proposed approach was implemented in VHDL on Altera Arria 10 GX FPGA. An energy-efficient design of computation system for deep neural networks with the usage of edge devices is presented in [9]. A novel hardware accelerator with low-precision computation and sparsity-aware structured zero-skipping, aimed to maximize energy efficiency designed, on top of the systolic-array structure described and prototype realized with the help of Xilinx Alveo U250 FPGA. Binary convolutional neural networks (BCNN) are present in [10], where bitwise operations are used instead of arithmetic operations, aimed to reduce required memory size, which is suitable to FPGAs to shorten training or inference. A low-end Zynq-7010 FPGA spike decoding system implementation is presented in [11], based on a multiplier-less spike detection pipeline with the help of a spiking neural network-based decoder constructed on programmable logic.

The paper [12] presents a table-algorithmic method for scalar product calculating of floating-point format operands and the structure of neural network based data cryptographic modules. Proposed decryption module implemented in Quartus II in VHDL language and its work simulation performed. Field-programmable gate arrays are used [13] to create neural networks in hardware (FPGAs). A larger neural network can be implemented in this case at a lower cost on a single chip. To realize fast execution for Homomorphic Encryption (HE), as a promising solution in case of increasing concerns of machine learning privacy, [14] proposed linear layers computations FPGA accelerator, to reduce the computational bottleneck in HE SCNN batch inference. As shown in [15] practical deployment of convolutional neural network (CNN) and cryptography algorithm on constrained devices are challenging due to the huge computation and memory requirement, so it is proposed a viable solution to this issue by expressing the CNN and cryptography as Generic-Matrix-Multiplication (GEMM) operations and map them to the same accelerator for reduced hardware consumption. FPGA-based lightweight cryptographic algorithms with the advantage of partial reconfiguration to secure designs are presented in [16]. Attempt at hardware architecture and implementation of NTRU prime system describer in [17]. A new scheme to secure Internet of Things data processing in public clouds concerning various attacks, especially from insiders, aimed at FPGA implementation, is proposed in [18]. So, as the analysis shows, the search for solutions for the FPGA-based implementation of neuro-like data cryptographic transmission is an actual task.

The aim of this work is the development of base components for neuro-like cryptographic implementation using polynomials for FPGA. For this, a structural diagram of the base encryption operation has been developed; the polynomial data encryption algorithm using tables of weighting coefficients for hardware implementation on FPGA is proposed. The preparation of the matrix of weight coefficients of the neural network is carried out by a program that, in addition, determines their maximum order and saves the obtained matrices in files.

The main preparatory stages of encryption/decryption of data based on NLN include:

- Selection of NLN architecture for encryption/decryption of data.
- NLN training, which consists in calculating weighting factors for ANN network for encryption/decryption data.
- Preparation of weighting coefficient arrays for the selected NLN architecture.

These stages are implemented in software and allow obtaining all the necessary data for the hardware implementation on the FPGA of NLN data encryption/decryption. Encryption can be made over text messages, control commands, or data. The secret key in proposed NLNs is a set of network architecture configurations and weighting coefficient arrays. In [19] described a method for parameter determination for the neural network architecture aimed at data encryption/decryption. The choice of the number of neuro elements and the number of inputs depends on the bit size of the message and the bit size of the inputs.

II. Neuro-like Data Encryption Algorithm Using Polynomials

The proposed data encryption algorithm is based on an ANN, which is trained without a teacher. For this purpose, a model of geometric transformations [6] is used. The architecture of such a neuro-like network consists of input, hidden, and output layers, and there are lateral connections between the neurons of the hidden layer. The outputs of the hidden layer neurons are connected to the inputs of NLN output layer neurons by polynomial connections. The number of neurons in all NLN layers is the same.

During encryption, the input data vectors are fed to the outputs of the neurons of the hidden layer. During decryption, the encrypted data is fed to the NLN inputs. We receive the decoded data at the hidden layer neurons output. The input data for NLN are positive integers: vectors of a certain bit size *n*, which are divided into *n_IN* vectors of bit size *k_IN.*

A graph model of a neuro-like data encryption algorithm has been presented in [19]. The proposed model provides a selection of the hardware structure to process different data intensities. Such a structure is focused on effective FPGA implementation of the base operation of nonlinear neuro-like encryption.

The basic data encryption operation is described by an expression.

$$y_j = \sum_{i=1}^{n\_IN} \sum_{p=1}^{P_{\max}} W_{ji}^{(p)} x_i^p \,, \tag{1}$$

where j=1, …, *n_IN*; *n_IN* – is the number of neurons in the input, hidden, and output layers of the NLN; $P_{\max}$ – is the maximum value of raising the input data to the power; x_i – input data with bit size *k_IN*, $W_{ji}^{(p)}$ – weighting coefficients, which are known values in floating point format and are obtained during NLN training.

The paper considers the algorithm of neural-like encryption of data with *n=16* bits using polynomials with

NLN architecture *n_IN=4* and *k_IN=4*, $P_{\max}$=3. As a result of NLN training, we get 4 matrices of weighting coefficients with dimension $P_{\max}$*n_IN, which are used in the encryption algorithm. Encrypted data is a vector of real numbers with dimension *n_IN*. One of the matrices of NLN weighting coefficients with *n_IN=4* and *k_IN=4* has the form:

$$
\begin{array}{llll}
W_1=[-6.497100 & -7.874508 & 7.193452 & 2.197609 \\
-2.383211 & -0.235815 & 0.086292 & 0 \\
3.454225 & 1.342763 & -0.216463 & 0].
\end{array}
$$

Let it be necessary to encrypt a vector of input data

$$X=[1100\ 0101\ 1001\ 1100],$$

which is divided into 4 input vectors of bit 4. These vectors are fed to the outputs of the hidden layer of the NLN:

$$X_1=[1100]_2 = 12,\ X_2=[0101]_2 = 5,$$
$$X_3=[1001]_2 = 9,\ \ X_4=[1110]_2 = 14.$$

Using expression (1) and the matrix of weighting factors W_1, we get Y_1=5615.026. Similarly, using matrices W_2, W_3, W_4, we will get encrypted data vectors Y_2=532.139, Y_3=1615.568, Y_4=322.879. So, vectors of input data X_1, X_2, X_3, X_4 correspond to vectors of encrypted data Y_1, Y_2, Y_3, Y_4.

To simplify the FPGA implementation of the neural-like data encryption algorithm using polynomials, the weighting coefficients are normalized and converted into 24-bit integers. The highest digit is a sign. Negative values are converted into a complementary code. This makes it possible to use built-in FPGA hardware multipliers of the Cyclone III family and megafunctions of the Quartus II library to implement the base encryption operation.

The transformation of the matrix of weighting coefficients from the real numbers format to the integers format consists of the following steps: determination of the maximum order $\max E_W$ among the elements of all matrices of weighting coefficients; calculation of the difference $\Delta E_{W_j} = \max E_W - E_{W_j}$ of orders for each of the elements of the matrices of weighting coefficients; shift to the right of each of the mantissa m_{W_j} elements of the matrix of weighting coefficients by its order difference ΔE_{W_j}; shift to the left of the scaled mantissa elements of the matrices of weight coefficients by 23 digits. For example, the value -6.497100 after transformations will take the value 0xCC05F1 (-3406351 in the decimal system), taking into account the maximum order among the elements of the matrices $\max E_W$=4. Really:

$$(-3406351/2^{23})*2^4=-6.497099.$$

After conversion to the integers format the weighting coefficients matrix W_1 will be following:

$$
\begin{array}{llll}
W_1=[-3406351 & -4128510 & 3771440 & 1152180 \\
-1249488 & -123635 & 45242 & 0 \\
1811008 & 703994 & -113489 & 0].
\end{array}
$$

III. Hardware Implementation of the Neuro-like Data Encryption Module Using Polynomials

The principle of operation of the FPGA-based neural-like data encryption module using polynomials is described in detail in [4]. Implementation of the module and its work simulation performed in the VHDL in the Quartus II v. 13.1 development environment based on the Cyclone III EP3C16F484C6 family FPGA [20]. In project components of the environment library [21] is used.

The main components of the neuro-like data encryption module using polynomials: block for exponentiation of the vector of input data X_i, ROM for storing tables of weighting coefficients, block for implementing the basic encryption operation. We will describe these components of the encryption module for the trained NLN with the resulting matrices of weight coefficients W_1, W_2, W_3, W_4.

The exponentiation component symbol is shown in Fig. 1, schematic diagram presented in Fig. 2.

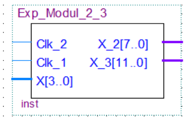


Fig. 1. Exponentiation component symbol

The Exp_Modul_2_3 component receives input data X[3..0] with k_IN=4 bits and synchronization pulses Clk_1, Clk_2. At the output of the components, we receive input data raised to the square (X_2[7..0]) and to the cube (X_3[11..0]). The megafunction ALTMULT_ADD[8] of the Quartus II development environment was used to implement this component.

The result of squaring off the input data by the leading edge of the *Clk*_1 pulses is written into the output register of the *Mult*_4*Bit* component. Similarly, the *Clk*_2 pulse synchronizes the recording of the cubed result of the *Mult*_12*Bit* component in the output register.

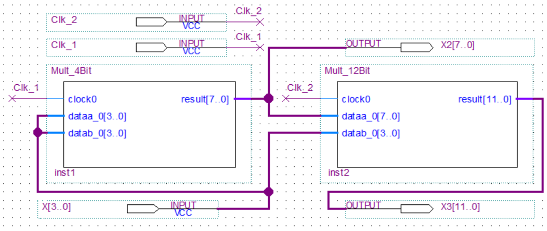


Fig. 2. Diagram of the exponentiation block

The delay time of the signal in the Exp_Modul_2_3 component is about 14 nsec. Hardware resources are required to implement the Exp_Modul_2_3 component: two 9-bit built-in multipliers, eight registers and eight logic elements. According to a similar scheme, the components of elevation to the 2nd (Exp_Modul_2) and 4th (Exp_Modul_2_3_4) powers of input data of 4 and 8 bits can be implemented.

The time diagrams of the simulation of the operation of the Exp_Modul_2_3 component are shown in Fig. 3.

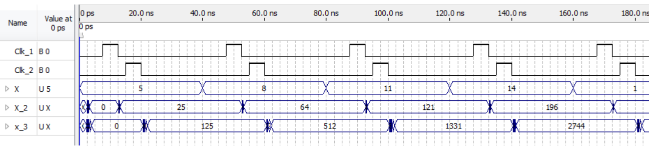


Fig. 3. Timing diagrams of the Exp_Modul_2_3 component simulation

The appearance of the symbol component for the base polynomial encryption operation is shown in Fig. 4, its schematics – in Fig. 5. The Mult_Add_4_3 component receives input data *X*_1[11..0], *X*_2[11..0] – input data raised to a square, *X*_3[11..0] – input data raised to a 12-bit cube (after equalization data bits *X*_1[3..0] and *X*_2[7..0]), weight coefficient vectors *W*_1[23..0], *W*_2[23..0], *W*_3[23..0], pulses timings *Clk*, *Clk*_1,, *Reset* pulse.

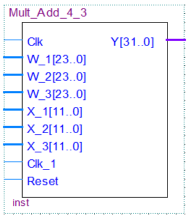


Fig. 4. Symbol component of the base encryption operation

At the component output value of the expression (1) is calculated *Y*[31..0]. For its implementation, megafunctions ALTMULT_ADD (component MULT_ADD_3), 4th registers of parallel type Reg_P_P, register R_Shift, megafunction PARALLEL_ADD (component ADD_Par) were used. Hardware resources used in the implementation of the Mult_Add_4_3 component: twelve 9-bit built-in multipliers, 375 logic elements, and 190 registers. The number of Mult_Add_4_3 components required for the input data encryption is determined by the *n*_*IN* value.

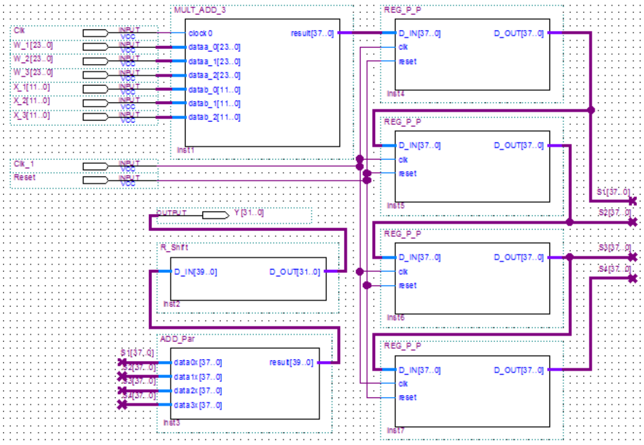


Fig. 5. Scheme of the components of the base encryption operation

The Table_W_Read_1 component of reading the weighting factors vectors of the W1 matrix from the FPGA ROM, squaring and cubing the input data vector, aligning the bitness of the vectors *X*[3..0], *X*_2[7..0]. A component is shown in Fig. 6, schematic diagram – in Fig. 7. These weighting factors are used only to calculate the Y_1 component of the encrypted data vector. Table_W_Read_2, …, Table_W_Read_4 components are used only for reading the remaining matrices of weighting factors W_2, W_3, and W_4.

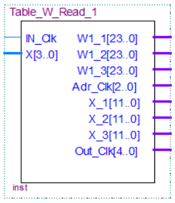


Fig. 6. Symbol of the components for reading of the weighting coefficients vectors

At the input of the Table_M_Read_1 component the input data vector X of 24 bits and the *IN_Clk* synchronization pulses are served. At the component output synchronization pulses *Out_Clk*[4..0] for the component of the basic encryption operation, vectors *X*_1, *X*_2, *X*_3, and vectors of weight coefficients *W*1_1, *W*1_2, *W*1_3 are forming. At the outputs of the Table_M_Read_2, ..., Table_M_Read_4 components, we get vectors of weight coefficients *W*2_1, *W*2_2, *W*2_3, …, *W*4_1, *W*4_2, *W*4_3. Synchronization pulses *IN_Clk* feed to the input of the *Clk_Control* and *C_Clk* components. After that at the *Clk_Control* component output, a vector of synchronization pulses *Out_Clk*[4..0] is forming, the values of which are read from the Clk_Control.mif table.

With the help of components W_ROM_1_4_1, W_ROM_1_4_2, W_ROM_1_4_3, the values of weight coefficients $W_{1,j}$ are read from the tables W_ROM_1_4_1.mif, W_ROM_1_4_2.mif, W_ROM_1_4_3.mif. The table cell address *Adr_Clk*[2..0] and the synchronization pulse *Out_Clk*[3], are fed to their input, on the leading edge of which data is read.

The data encryption schematics are shown in Fig. 8, and results of the simulation are shown in time diagrams (Fig. 9).

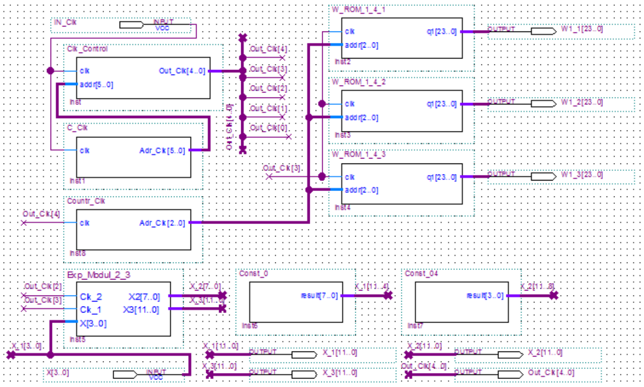


Fig. 7. Scheme of the Table_W_Read_1 component for reading vectors of weighting coefficients

The schematic diagram of the input data vector encryption for *n*_*IN*=4 and *k*_*IN*=4 is shown in Fig. 8. The hardware resources required for its implementation are 50 (out of 112) 9-bit built-in multipliers, 1480 (10%) logic elements, 794 (5%) registers. Time diagrams (Fig. 9) show us that the calculation of Y_1, Y_2, Y_3*, and* Y_4 based on expression (1) with *n*_*IN*=4 with their parallel implementation takes about 180 ns. In rows *Y*_1, *Y*_2, *Y*_3, *Y*_4 (Fig. 9) we received the results of encryption: values 11499569, 1089819, 3308677, 6612568 in the format of 32 bits integers.

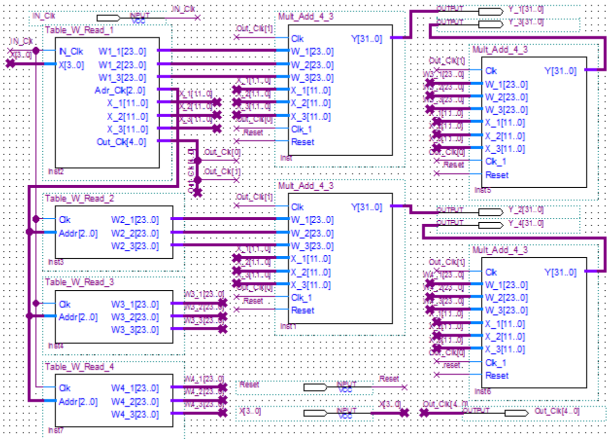


Fig. 8. Scheme of the encryption module

To convert from the format of integers to the format of real numbers, and perform their decryption, you need to know the maximum order maxE_W of the matrices of weighting coefficients for encryption and decryption – an additional secret key that does not need to be transmitted with the encrypted data (in our case, 2^4):

$11499569/2^4/2^7=5615.024;$
$1089819/2^4/2^7 =532.138;$
$3308677/2^4/2^7 =1615.565;$
$6612568/2^4/2^7 =3228.793.$

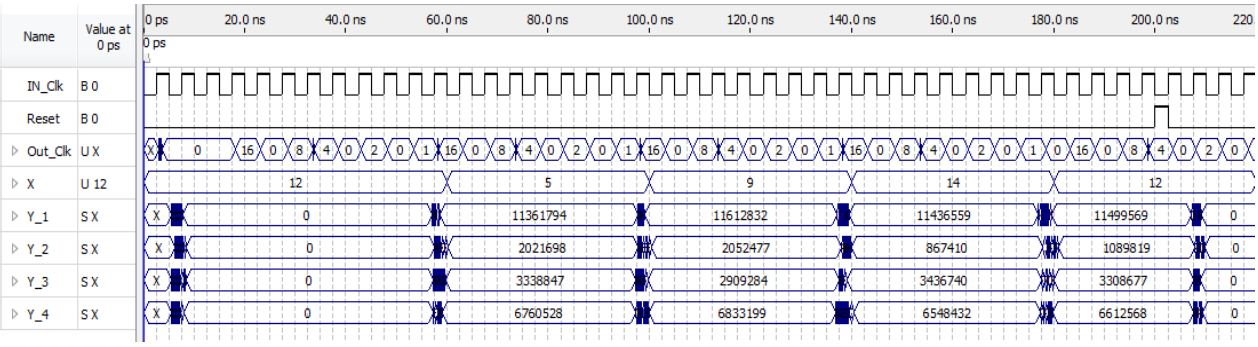


Fig. 9. Time diagrams of data encryption simulation

These results agree with the Y_1, Y_2, Y_3, Y_4 values obtained using calculations.

IV. Data Preparation for Encryption Algorithm Hardware Implemention

For the preparation of matrices of weighting coefficients and input data in the C language implemented Training_NS_MGT program and the Transfor_Data program for training NLN based on the model of geometric transformations. The Training_NS_MGT program, using generated random input data of a given bit rate, calculates the matrix of coefficients $a_{pi}^{(m)}$, (i=0, ..., n_{IN}-1; m=1, …, n_{IN}, p=1, …, $P_{\max}$), which is used for encryption of input data and matrix of weighting coefficients $M_i^{(m)}$ (i=0, ..., n_{IN}-1; m=1, …,n_{IN}) for a given NLN architecture.

The Transfor_Data program reads files with NLN architecture and matrices $a_{pi}^{(m)}$, $M_i^{(m)}$. From the matrix of coefficients $a_{pi}^{(m)}$, the matrices of weighting coefficients W_1, …, $W_{\mathrm{n\_IN}}$, are formed, the maximum order is found among the elements of the matrices $M_i^{(m)}$ and $a_{pi}^{(m)}$, the received weighting coefficients matrices are converted from the real numbers format to the integers format and stored in files.

For the Training_NS_MGT program to work, you need to set the NLN architecture: N – the number of training vectors; n – bit rate of training vectors; *k_IN* is the bit rate of neurons of the NLN input layer; *n_IN* is the number of neurons of the NLN input layer. Also, *X_Start* – the initial value for generating a sequence of pseudorandom numbers, is entered in the program.

To train NLN based on MGT, the Training_NS_MGT program uses the matrix of input data X[N, *n_IN*], formed by pseudo-random numbers in the [0,…, $2^{k_{IN}} -1$] range. The *k_IN* value affects the maximum order of matrix elements $a_{pi}^{(m)}$, $M_i^{(m)}$ For comparison, the matrix of weighting coefficients W_1 for NLN with architecture n=32, *n_IN*=4, *k_IN*=8 with the same value of $P_{\max}$=3 is given.

```
W1=[88.45100 -21.65212 -141.76999 -70.38724
     9.31598  -1.85190    7.33200    0
    24.24050 -11.19896   -8.83328    0].
```

Conclusion

Features of NLN based on MGT are its fast non-iterative learning for a predetermined number of calculation steps and repeatability of learning results. The key elements in neuro-like data encryption using polynomials are the architecture and parameters of the NLN after its training. Training_NS_MGT and Transfor_Data programs are written in C for the Raspberry Pi 4 microcomputer. The performance of this microcomputer allows it to effectively perform training of the NLN and encryption/decryption of data and work as part of mobile systems. The Training_NS_MGT program trains NLN with a given architecture based on MGT, and forms matrices used in neuro-like data encryption using polynomials. The Transfor_Data program prepares the parameters, obtained during NLN training, for their use in the encryption algorithm in the FPGA hardware implementation. The Mult_Add_4_3 component is developed in the VHDL with the help of Quartus II ver. 13.1 development environment for Cyclone III EP3C16F484C6 FPGA. It is implemented with FPGA hardware multipliers and megafunctions MULTALT_ADD, PARALLEL_ADD and corresponds to the basic operation of neural-like data encryption using polynomials based on NPM with *n_IN*=4 and *k_IN*=4.

The scheme of the neuro-like encryption module of input data with *n*=16 bits using the Mult_Add_4_3 component is implemented on FPGA. The hardware resources necessary for its implementation are given. A library of similarly implemented base components for other NLN architectures will allow them to be effectively used to implement neuro-like encryption modules with different bit widths of input data. The simulation of the operation of the circuit of the neuro-like encryption module was performed. The obtained time diagrams (Y_1, Y_2, Y_3, Y_4) confirm the correctness of its operation. The encryption time for *n*=4, *n_IN*=4, *k_IN*=4 is about 180 ns and depends most significantly on *n_IN*. As the number of NLN entries increases, the encryption time will increase. The hardware resources needed to implement the scheme of the neuro-like cryptographic data protection module will also increase. The direction of further research is the hardware implementation of the neural-like decoding algorithm on FPGA in the form of library components and the decoding module, testing its performance.

References

[1] A. El-Zoghabi, A.H. Yassin and H.H. Hussien, "Survey report on cryptography based on neural network", International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 12, pp. 456-462, 2013.

[2] Karam M. Z. Othman, Mohammed H. AL Jammas, "Implementation of neural - cryptographic system using FPGA", Journal of Engineering Science and Technology, Vol. 6, No. 4, 411 – 428, 2011.

[3] I. Tsmots, V. Rabyk, Yu. Lukaschuk, V. Teslyuk, Z. Liubun "Neural Network Technology for Protecting Cryptographic Data", in: Proceedings of the XIIth International Scientific and Practical Conference "Electronics and Information Technologies". 2021. P. 103 – 108.

[4] I. Tsmots, R. Tkachenko, V. Teslyuk, Y. Opotyak and V. Rabyk, "Hardware Components for Nonlinear Neuro-like Data Protection in Mobile Smart Systems", in: IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 2022, pp. 198-202, doi: 10.1109/CSIT56902.2022.10000636.

[5] Diamantaras K. I., Kung S. Y. Principal Component Neural Networks: Theory and Applications, John Wiley & Sons, 1996. 272 p.

[6] I. Izonin, R. Tkachenko, N. Kryvinska, P. Tkachenko, and M. Greguš ml., 'Multiple Linear Regression Based on Coefficients Identification Using Non-iterative SGTM Neural-like Structure', in Advances in Computational Intelligence, I. Rojas, G. Joya, and A. Catala, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 467–479. doi: 10.1007/978-3-030-20521-8_39.

[7] S. Feizi, A. Nemati, S. Haghiri, A. Ahmadi and M. Seif, "Digital Hardware Implementation of Lightweight Cryptography Algorithm Using Neural Networks", in: 28th Iranian Conference on Electrical Engineering (ICEE), Tabriz, Iran, 2020, pp. 1-7, doi: 10.1109/ICEE50131.2020.9260614.

[8] K. Khalil, O. Eldash, B. Dey, A. Kumar and M. Bayoumi, "Architecture of A Novel Low-Cost Hardware Neural Network", in: IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS), Springfield, MA, USA, 2020, pp. 1060-1063, doi: 10.1109/MWSCAS48704.2020.9184585.

[9] S. Kim, S. Cho, E. Park and S. Yoo, "FPGA Prototyping of Systolic Array-based Accelerator for Low-Precision Inference of Deep Neural Networks", in: 2021 IEEE International Workshop on Rapid System Prototyping (RSP), Paris, France, 2021, pp. 1-7, doi: 10.1109/RSP53691.2021.9806200.

[10] S. Kim and R. Rutenbar, "Accelerator Design with Effective Resource Utilization for Binary Convolutional Neural Networks on an FPGA", in: IEEE 26th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), Boulder, CO, USA, 2018, pp. 218-218, doi: 10.1109/FCCM.2018.00052.

[11] G. Leone, L. Raffo and P. Meloni, "On-FPGA Spiking Neural Networks for End-to-End Neural Decoding," in IEEE Access, vol. 11, pp. 41387-41399, 2023, doi: 10.1109/ACCESS.2023.3269598.

[12] I. Tsmots, V. Rabyk, V. Teslyuk and Y. Opotyak, "Floating-Point Number Scalar Product Hardware Implementation for Embedded Systems", in: 17th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), Jaroslaw, Poland, 2023, pp. 6-10, doi: 10.1109/CADSM58174.2023.10076502.

[13] M. T. Ali and B. H. Abd, "An Efficient area Neural Network Implementation using tan-sigmoid Look up Table Method Based on FPGA", in: 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-7, doi: 10.1109/INCET54531.2022.9825348.

[14] Y. Yang, S. R. Kuppannagari, R. Kannan and V. K. Prasanna, "FPGA Accelerator for Homomorphic Encrypted Sparse Convolutional Neural Network Inference", in: IEEE 30th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), New York City, NY, USA, 2022, pp. 1-9, doi: 10.1109/FCCM53951.2022.9786115.

[15] J. -C. See. H.-F. Ng, H.-K. Tan, J.-J. Chang, K.-M. Mok "Cryptensor: A Resource-Shared Co-Processor to Accelerate Convolutional Neural Network and Polynomial Convolution", in: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, doi: 10.1109/TCAD.2023.3296375.

[16] F. G. Bîrleanu and N. Bizon, "Lightweight cryptography for Internet of Things using FPGA-based Design with Partial Reconfiguration", in: 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Bucharest, Romania, 2020, pp. 1-7, doi: 10.1109/ECAI50035.2020.9223213.

[17] H. Wu and X. Gao, "Efficient Multiplier and FPGA Implementation for NTRU Prime", in: IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), ON, Canada, 2021, pp. 1-5, doi: 10.1109/CCECE53047.2021.9569160.

[18] M. Al-Asli, M. E. S. Elrabaa and M. Abu-Amara, "FPGA-Based Symmetric Re-Encryption Scheme to Secure Data Processing for Cloud-Integrated Internet of Things", in: IEEE Internet of Things Journal, vol. 6, no. 1, pp. 446-457, Feb. 2019, doi: 10.1109/JIOT.2018.2864513.

[19] Tsmots, I., Teslyuk, V., Lukashchuk, Y., Opotiak, Y. "Method of Training and Implementation on the Basis of Neural Networks of Cryptographic Data Protection", in: CEUR Workshop Proceedings, 2022, 3171, pp. 916–928.

[20] Cyclone III Device Handbook, 2023. URL: https://cdrdv2-public.intel.com/654357/cyclone3_handbook.pdf

[21] Integer Arithmetic IP Cores User Guide, 2023. URL: https://vanhunteradams.com/DE1/Drum/ug_lpm_alt_mfug.pdf

An Approach for Automated Code Deployment between Multiple Node.js Microservices

Oleh Chaplia
Department of Specialized Computer Systems
Lviv Polytechnic National University
Lviv, Ukraine
oleh.y.chaplia@lpnu.ua

Halyna Klym
Department of Specialized Computer Systems
Lviv Polytechnic National University
Lviv, Ukraine
halyna.i.klym@lpnu.ua

***Abstract* — Microservices represent an architectural approach in software development, where a complex solution is formed by combining small, independently deployable services that communicate through specified interfaces (APIs). These microservices have diverse applications across sectors such as banking, finance, healthcare, e-commerce, government, military, logistics, and gaming. The microservices architecture offers fundamental guidelines on how these small services should function as part of a more prominent solution.**

The research introduces a framework based on Node.js for facilitating automated sharing of source code among services and their execution within a distributed system environment. This proposed framework, constructed on the Node.js platform, features a compact codebase and mechanisms designed to enable sharing of user-developed code among multiple microservices instances within a single network.

The selection of Node.js as the foundation for this framework is motivated by its compact nature and suitability for building server applications. Furthermore, Node.js can operate effectively on various devices, including embedded systems, mobile phones, and devices with limited operating system resources.

Keywords* — *node.js; microservices; cloud architecture; distributed systems; framework.

I. Introduction

Microservices, an architectural approach in software development, involve constructing complex solutions from independently deployable services communicating through specific APIs [1, 2]. These services cater to necessary business functionalities. Particularly beneficial for sizable and intricate applications, microservices offer advantages like flexibility, scalability, and accelerated development [2].

However, implementing microservices demands careful planning, seamless interaction, precise coordination, and a clearly outlined project structure. Software engineers and developers must meticulously account for these aspects during microservices development [1, 2, 3].

Node.js is a framework and library for constructing server applications using JavaScript [4]. Its suitability for microservices and server applications stems from its efficient handling of asynchronous and non-blocking I/O operations, its lightweight runtime, scalability, and the extensive JavaScript ecosystem it offers [4]. Additionally, Node.js excels in event-driven architectures and high-performance solutions [4].

JavaScript follows a prototype-based approach, allowing developers to employ functional or object-oriented programming paradigms [5]. For those seeking stricter type enforcement, TypeScript can be utilized [5].

One of the challenges encountered with multiple microservices [2] involves managing the codebase. Notably, each microservice repository often contains duplications of core libraries such as HTTP services, loggers, configuration managers, API controllers, and database connectors [2]. This code redundancy is not uncommon [2].

Scaling and deployment issues [2, 6, 7] are another consideration. Node.js applications are frequently scaled using Docker containers and Kubernetes [6, 8, 9]. Kubernetes achieves scaling by duplicating instances and effectively managing client requests across machines. Most major cloud providers, like AWS and Microsoft Azure, offer automated scaling tools for Node.js applications [2, 6]. These approaches entail replicating identical containers with the same source code and employing a load balancer (connection manager) to handle requests and responses [7, 10].

Kubernetes or similar tools may not be available or suitable in specific situations. For instance, they deploy the same microservice solution in on-premises data centers or on small, embedded devices connected through a network. In such cases, the solution's footprint must be minimal. These compact, embedded devices serve purposes like data collection, monitoring, healthcare, or military applications.

II. Approaches for Scaling Node.js Microservices

Horizontal scaling and vertical scaling are two distinct approaches to improving the capacity and performance of a system or application [2, 10].

Horizontal scaling, also known as scaling out, involves adding more machines or instances to a system. This means that multiple copies of the application run on separate machines, and incoming requests are spread across these instances. This approach optimizes the system's ability to handle increased traffic and a heavier workload [2, 10].

On the other hand, vertical scaling, or scaling up, involves enhancing the resources of a single machine or instance, such as CPU, memory, or storage. This is typically done through hardware upgrades or more powerful virtual machines [2, 10].

The choice between horizontal and vertical scaling depends on various factors, including the architecture of the application, expected growth patterns, resource availability, and cost considerations. In some cases, hybrid strategies that combine horizontal and vertical scaling methods are used to balance performance, availability, and manageability [2, 10].

In the context of efficiently expanding Node.js microservices [8], Docker and Kubernetes play essential roles. Docker simplifies the creation of lightweight and self-contained containers. When there's a need to scale up, Docker

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

makes it easy to duplicate containers to handle increased demand [8]. On the other hand, Kubernetes offers automated container management [8]. It effectively takes tasks like deploying, expanding, and managing container-based applications. Kubernetes' auto-scaling feature dynamically adjusts the number of containers instances and addresses the distribution of incoming traffic among multiple microservice instances to prevent overloading any single instance. In summary, Docker's containerization simplifies Node.js microservice deployment, while Kubernetes automates orchestration and scaling procedures [8]. This powerful combination ensures that microservices can adapt to changing demands while maintaining consistent performance.

Prominent cloud providers like AWS, Microsoft Azure, and IBM Cloud offer tools and methods for scaling Node.js applications. For instance, AWS provides various tools designed for scaling [10, 11]. One such tool is Elastic Load Balancing (ELB), engineered to evenly distribute incoming traffic among multiple instances of a Node.js application. This prevents any instance from becoming overwhelmed and adapts dynamically based on traffic patterns [11]. AWS Auto Scaling empowers developers to set up policies for automatic adjustments to the number of instances in response to resource usage or custom metrics [11]. Amazon EC2 Instances offer adjustable computing capacity, allowing developers to choose instances tailored to the specific CPU, memory, and storage requirements of their Node.js applications [11]. Amazon ECS streamlines the management of Node.js applications within Docker containers, enhancing flexibility and scalability [11]. AWS Lambda supports serverless architectures by automatically adjusting the execution of Node.js code in response to events [11].

Similarly, Microsoft Azure, a prominent cloud service provider, presents a range of tools for effectively scaling Node.js applications [2, 12]. Azure's Elastic Load Balancing services, such as Azure Load Balancer and Application Gateway, distribute incoming traffic evenly among multiple Node.js instances [12]. The Auto Scaling feature dynamically adjusts the number of active Node.js instances in response to real-time demand [2, 10, 12], ensuring seamless adaptation to fluctuating traffic without requiring manual intervention [12]. Azure App Service simplifies Node.js application deployment by handling infrastructure setup, load distribution, and scaling. This allows developers to focus on application development rather than infrastructure management [2, 10, 12]. For more complex scenarios, Azure Kubernetes Service (AKS) offers managed Kubernetes clusters, enabling developers to efficiently manage and scale containerized Node.js applications using Kubernetes' advanced capabilities [8].

Despite the diverse scaling methods and tools, they all have a common feature: they replicate a single instance of Node.js microservices into multiple copies specified by the user and handle network connections among them. Nevertheless, they do not scale the source code while the program runs. This research puts forward a solution to tackle this challenge.

This research presents an innovative approach to maintaining a single source code base within a microservice while automating the distribution and execution of this code across multiple services. This approach is designed explicitly for Node.js. The framework's architecture for Node.js is consistent and spans all deployable services. It supports modular design, initializes all deployed services, and automates code scaling and execution across chosen microservices. Additionally, the provided architecture aims to be concise and easy for developers and engineers to understand. The central focus of this research centers on a Node.js framework created to scale user application code across multiple microservices and execute it accurately, all while automatically handling dependencies among microservices.

III. NODE.JS FRAMEWORK SOLUTION

The proposed Node.js framework is built upon a model that employs a primary and replica approach. An overview of the framework's architecture is illustrated in Fig. 1. The primary microservice serves as the central hub for housing the business logic code and the framework for code scaling and execution. The replica microservices are deployed instances that run a copy of the framework but do not contain any business logic code. Both primary and replica microservices are flexible enough to be deployed on various cloud providers like AWS or Azure or within an on-premise network of IoT devices. This flexibility allows developers to choose the deployment environment that suits their needs.

Each primary and replica microservice should be accessible externally through external IP or DNS addresses. It's important to note that the primary microservice must be aware of the IP addresses of all replica microservices. The source code for the primary microservice should be consolidated within a single project, including the source code for each microservice intended for deployment. For example, the code repository can be hosted on a platform such as GitHub, and deployment is restricted to a single instance of the primary microservice.

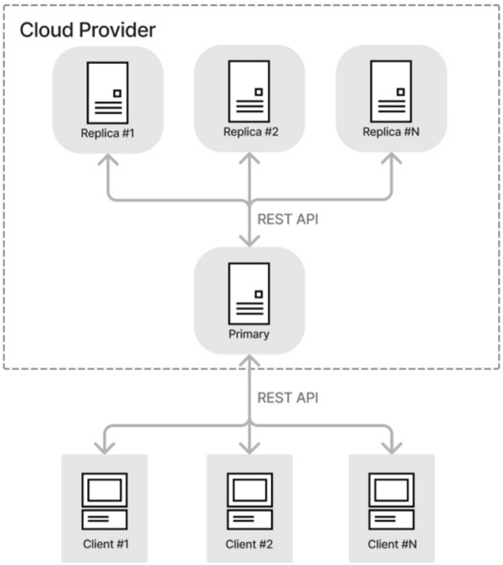


Fig. 1. Architecture of a Node.js framework

After transferring the source code from a Git repository to a primary microservice instance and completing the build process, the initialization phase begins. Fig. 2 provides an overview of initializing the primary microservice and handling client requests. First, the primary microservice compiles a comprehensive list of addresses for all deployed services and verifies their availability by ping each address. Next, the framework automatically distributes the user code to all replica microservices, requiring no manual intervention. Once this code is disseminated, the primary microservice receives notification that all replicas are ready to operate. After next of the user code, the replicas do not need to be redeployed.

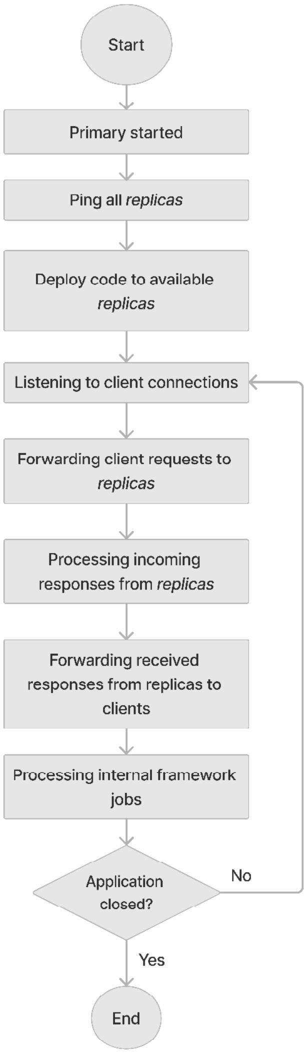

Start
Primary started
Ping all *replicas*
Deploy code to available *replicas*
Listening to client connections
Forwarding client requests to *replicas*
Processing incoming responses from *replicas*
Forwarding received responses from replicas to clients
Processing internal framework jobs
Application closed?
No
Yes
End

Fig. 2. Processing user requests by a primary microservice.

The primary microservice is a central hub responsible for processing all client requests. Behind the scenes, the framework efficiently manages the execution of user code, which includes the business logic. It also initiates internal calls to the replica microservices using preconfigured addresses, gathers the results, and sends responses back to the clients.

From the client's perspective, they don't need to be aware of the existence of replica microservices. Clients interact with the system by invoking public REST APIs or using various communication protocols to communicate with the primary microservice. They send their requests to the primary microservice and wait for responses.

The storage repository for user code, which contains the essential business logic, is located within the Node.js framework ecosystem of the primary microservice. Developers can create user code according to their preferences and specific requirements.

IV. Deployment and Build Process of the Proposed Node.js Framework

The build process consists of two separate stages. The first step involves deploying all the replica microservices and the primary microservice code. In this example, the source codes for both the primary and replica microservices are stored in two separate GitHub repositories. These GitHub repositories are connected seamlessly to their respective machine instances through cloud provider tools. The processes for deploying the primary and replica microservices are illustrated in Fig. 3 and Fig. 4.

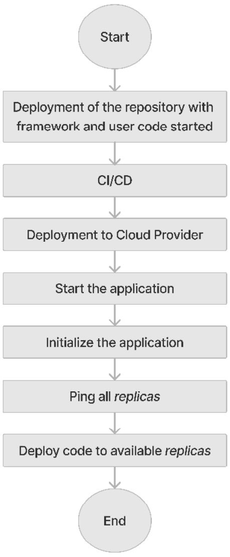

Start
Deployment of the repository with framework and user code started
CI/CD
Deployment to Cloud Provider
Start the application
Initialize the application
Ping all *replicas*
Deploy code to available *replicas*
End

Fig. 3. The process of the deployment of primary microservice

The deployment process follows a straightforward sequence. When developers create a pull/merge request from the development branch to the main branch, they review and merge it. After the merge, the developer's defined CI/CD (Continuous Integration/Continuous Deployment) pipeline is triggered. Subsequently, an event hook is sent to the cloud provider, instructing it to start loading the source code from the GitHub repository. Once the code is successfully loaded, the build and initialization process begin. It's crucial at this point to have the IP/DNS addresses of all replicas known. During the initialization phase, all replicas are validated through ping events. After successfully validating the replicas, the primary server is initiated and ready to handle user requests.

A test project was initiated to evaluate the proposed Node.js framework and analyze the outcomes. The current tests primarily focus on assessing the concept of primary and replica deployment, the distribution of user code, and the communication processes between the primary and replica microservices. It's worth noting that these tests do not cover aspects like load testing, network latency evaluation, or data transfer speed assessments.

Two GitHub repositories were set up for these tests, designated for the primary and replica microservices. The initial test suite includes one primary microservice and three replica microservices. The initial tests evaluated the primary and replica microservices locally, running on the same machine but on different ports. The primary and replica microservices were deployed to the cloud service provider Heroku in a subsequent test.

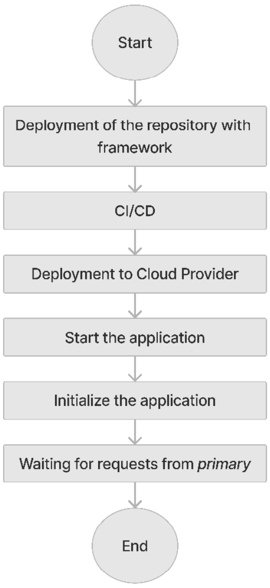

Start
Deployment of the repository with framework
CI/CD
Deployment to Cloud Provider
Start the application
Initialize the application
Waiting for requests from *primary*
End

Fig. 4. Process of the deployment of replica microservice

The primary objective of the test project was to confirm the validity of the architectural framework. This included evaluating the connections between various services and determining whether the framework could independently distribute and execute code for each microservice. While the project did not involve the implementation of specific business logic, interactions between the services were included to assess their communication capabilities.

V. Results

Results were collected for further analysis after successfully implementing the proof-of-concept and conducting subsequent testing. The initial test involved running both the primary and replica microservices locally. One primary microservice and three replica microservices were successfully initiated. All three replica microservices were pinged during the initialization process to confirm their availability. Since all of them were on the same machine, they were easily accessible. Following this, the primary microservice seamlessly shared the user code with each initialized replica microservice, ensuring that all replicas were prepared for code execution.

A series of three sequential test requests were sent to the primary microservice to evaluate the system's functionality. The first request invoked the user code from the first replica microservice. The primary microservice forwarded this request to the first replica microservice, waited for the response, returned the result to the primary microservice, and finally relayed it to the client. The second and third requests followed a similar pattern, each calling the user code from the second and third replicas.

Subsequent tests explored the parallel execution of identical requests. The behavior remained consistent with the earlier sequential tests, except that the test requests were processed simultaneously. Similar behavior patterns were observed when the microservices were deployed, albeit with variations in the addresses of the primary and replica microservices. Additionally, an increase in network latency was noted due to deploying the microservices to the cloud.

Conclusion

The development and validation of a Node.js framework designed to simplify code sharing and automate execution across multiple distributed machines have been successfully achieved. This proposed approach represents that developers can streamline their work processes by maintaining a single codebase containing all the essential business logic. It eliminates the need to manage separate repositories for each microservice. This approach significantly reduces the complexity of overseeing source code for various microservices within a distributed system. Moreover, the proposed Node.js framework enhances efficiency by automating the processes of code sharing and execution among microservices. Developers now have the flexibility to choose the number of replica microservices required, with the option for horizontal scaling also available.

To further advance this research, future investigations could explore optimization opportunities for communication protocols between microservices. Additionally, examining strategies for managing solution architectures featuring multiple primary microservices connected through load-balancing mechanisms could be beneficial. Further refinements of the automated code scaling and execution processes across a network of distributed machines could also be explored. These potential areas for research and improvement hold promise for enhancing the capabilities and adaptability of the proposed Node.js framework.

References

[1] G. Blinowski, A. Ojdowska, and A. Przybyłek, “Monolithic vs. Microservice Architecture: A Performance and Scalability Evaluation,” *IEEE Access,* vol. 10, pp. 20357-20374, 2022.

[2] A. S. Abdelfattah, and T. Cerny, “Roadmap to Reasoning in Microservice Systems: A Rapid Review,” Appl. Sci., vol. 13, no. 2, pp. 1838, 2023.

[3] B. Basumatary, N. Agnihotri, ”Benefits and Challenges of Using NodeJS,” *International Journal of Innovative Research in Computer Science and Technology (IJIRCST)*, vol. 10, no. 3, pp. 67-70, 2022.

[4] M. E. Gortney *et al*., “Visualizing Microservice Architecture in the Dynamic Perspective: A Systematic Mapping Study,” *IEEE Access*, vol. 10, pp. 119999-120012, 2022.

[5] M. Golec, M. Plechawska-Wójcik, “Comparative analysis of frameworks using TypeScript to build server applications,” *Journal of Computer Sciences Institute*, vol. 23, pp. 128-134, 2022.

[6] Zhu, J., Patros, P., Kent, K. B., and Dawson, M., “Node.js scalability investigation in the cloud,” pp. 201–212, 2018, Proceeding of 28th Annual International Conference on Computer Science and Software Engineering (CASCON 2018), New York, NY, USA: ACM.

[7] F. Doglio, Scaling Your Node.js Apps. Apress Berkeley, CA, 2018.

[8] K. Juell, From containers to Kubernetes with Node.js, 2020, New York City, New York, USA.

[9] T. Hunter II, Distributed Systems with Node.js: Building Enterprise-ready Backend Services, 2020, O'Reilly Media, Inc.

[10] P. Singh, P. Gupta, K. Jyoti, and A. Nayyar, “Research on Auto-Scaling of Web Applications in Cloud: Survey, Trends and Future Directions,” *Scalable Computing: Practice and Experience*, vol. 20, no. 2, pp. 399-431, 2019.

[11] O. Mustafa, A Complete Guide to DevOps with AWS, Apress, Berkeley, CA, 2023.

[12] D. DeJonghe, A. Nugara, Application Delivery and Load Balancing in Microsoft Azure, O'Reilly Media, Inc., 2020.

Identity Verification and Detection System through Facial Recognition Technology

Nazar Karpiuk
Department of Specialized Computer Systems
Lviv Polytechnic National University
Lviv, Ukraine
nazar.karpiuk.mkisk.2022@lpnu.ua

Halyna Klym
Department of Specialized Computer Systems
Lviv Polytechnic National University
Lviv, Ukraine
halyna.i.klym@lpnu.ua

***Abstract* — This investigation focuses on the examination and experimentation of the Haar cascade classifier technique for facial recognition. The study additionally rationalizes the selection of the algorithm and furnishes an account of how the system's implementation, utilized for the analysis, was carried out. Throughout the study, the developed system underwent testing using a set of facial photographs that encompassed a variety of attributes, including differing distances from the camera, diverse lighting conditions, and varied facial orientations within the camera's field of view. A thorough evaluation of the test outcomes was performed, leading to conclusions that shed light on the essential considerations in the design and application of facial recognition systems, aiming for optimal accuracy.**

Keywords* — *face recognition, Haar classifiers, eigenfaces, fisherfaces

I. Introduction

Advancements in the realms of automation, robotics, and artificial intelligence have ushered in an era of unprecedented convenience in human life. However, this technological progress has also been accompanied by an increasingly hectic lifestyle, leading to a heightened interest in security-focused technologies. Individuals across the globe are now utilizing a diverse array of security systems to safeguard their homes, vehicles, workplaces, mobile devices, laptops, and more. These security systems leverage a plethora of cutting-edge technologies, including remote keyless access, near-field communication, touchscreens, as well as biometric or multi-biometric systems. Notably, biometric systems stand out as the pinnacle of efficiency and accuracy, offering an elevated level of security [1].

Within the realm of biometric systems, the automatic recognition and detection of faces have emerged as a pivotal area of research and development. This innovative technology draws upon the principles of computer vision and image recognition, enabling the identification of individuals by comparing facial features with an extensive database of known faces. Beyond its primary role in security, facial recognition technology finds application in an array of domains, such as user authentication, criminal investigations, video surveillance, robotics, and the medical sciences [2].

Facial recognition technology has seamlessly integrated itself into modern applications, spanning from access control and security systems to personalized user experiences [3,4]. Among the various techniques employed for face detection, the Haar cascade classifier method has garnered widespread acclaim for its remarkable efficiency and accuracy [5,6]. In this scientific endeavor, we present the implementation of a facial recognition system founded upon the Haar cascade classifier method, executed on the versatile Arduino platform.

The convergence of facial recognition technology with Arduino, a widely accessible microcontroller board, ushers in an era of new possibilities for real-time and embedded face detection applications. Traditionally, facial recognition systems demanded substantial computational resources, but recent strides in hardware capabilities have rendered it feasible to deploy these systems on low-power microcontrollers like Arduino.

The initial phase of real-time automatic face recognition and detection unfolds with the aid of a camera that captures an image. This image is subsequently subjected to a multi-step process that includes face detection, identification of regions of interest (specific areas within the image where operations will be conducted), feature extraction, and ultimately, the recognition of faces.

Amidst the landscape of facial recognition and tracking systems, there exists a vibrant realm of research and development, where scientists from diverse corners of the world continually strive to devise, refine, and improve upon various approaches and algorithms. While machine learning models [7] and neural networks have achieved prominence within these systems, there exists a tapestry of alternative methodologies that remain under active exploration. Beyond the Haar cascade classifier method, which is the focal point of this scientific investigation, several other distinguished techniques warrant mention. These include Eigenfaces, Fisherfaces, and Local Binary Patterns Histograms (LBPH) [8]. Each of these methods exhibits distinct advantages and limitations, necessitating a careful consideration of their characteristics when devising specific facial recognition and tracking systems.

In essence, this domain exerts an ever-increasing pull on research resources, year by year, with the aim of refining existing algorithms and fostering the emergence of entirely novel approaches. As technology continues to evolve, the potential for innovation in the realm of facial recognition and tracking systems remains limitless, offering a compelling arena for scientific exploration and development.

II. Methodology

The utilization of the Haar cascade classifier method as the foundation for our face recognition and tracking system has proven to be a highly efficient technique for the detection of faces within images. Central to this methodology was the employment of the pre-trained standard classifier, haarcascade_frontalface_default.xml, an integral component of the OpenCV (ver. 4.5.5) software package. This classifier is endowed with a meticulously crafted set of rules designed expressly for the purpose of face detection, rendering it an ideal choice for our study.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

The system's prowess was rigorously tested by subjecting it to a battery of real-world scenarios, involving photographs captured from a webcam under varying conditions. These conditions encompassed different distances, lighting conditions, and an array of parameters, including photographs taken from diverse angles. The resultant dataset, replete with this extensive variety, served as the litmus test for evaluating the system's accuracy and robustness across a spectrum of challenging scenarios.

To quantify the system's performance, we employed the "face detection accuracy" metric. This metric, defined as the ratio of correctly detected faces to the total number of images within the test set, provided a precise quantitative gauge of the system's efficacy. Each measurement entailed the capture of 100 photographs, with the corresponding ratio meticulously calculated.

The code implementation was carried out using the Python programming language and the OpenCV library. Given that Python was chosen as the programming language, the system's primary computational requirement included the installation of Python 3.x, along with the essential libraries such as NumPy. To ensure optimal performance, it is recommended to have a multicore processor with a clock speed of at least 2.0 GHz or higher. Therefore, the computational requirements include sufficient RAM, with a minimum of 4 GB recommended for moderate image sizes

Ultimately, the "face detection accuracy" metric emerged as the linchpin of our evaluation framework, affording a precise and quantitative measure of the system's effectiveness. This meticulous approach to testing and measurement underscored the success of our system and provided valuable insights for its further refinement and optimization.

III. Prepare Your Paper Before Styling

The sheer diversity of functionality within facial recognition systems underscores the inherent complexity of addressing users' distinct needs. Nevertheless, by examining common features and the fundamental nature of these systems, it becomes possible to distill a generalized algorithm that serves as the foundation for their operation. Stripping away peripheral functions such as gesture recognition or autofocus, the entire process of facial recognition and tracking can be distilled into a sequential framework, as illustrated in Fig 1.

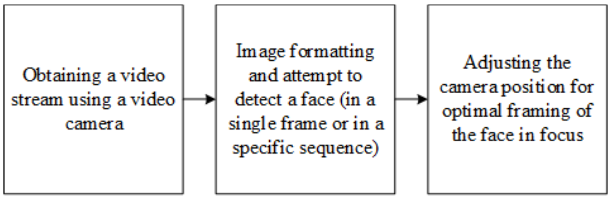

Obtaining a video stream using a video camera
Image formatting and attempt to detect a face (in a single frame or in a specific sequence)
Adjusting the camera position for optimal framing of the face in focus

Fig. 1. The sequence of events in the face recognition process

Among the myriad methods available for facial recognition and tracking, the Haar cascade classifier method stands out as one of the most effective. Particularly for an Arduino-based system that relies on a webcam for real-time image acquisition, the Haar method proves to be the most fitting choice.

The Haar method operates by distinguishing between primary and secondary image features. Primary features encapsulate images of the objects to be detected, while secondary features represent combinations of primary features designed to enhance object retrieval accuracy [6]. To harness the power of the Haar method in object detection, an initial training phase is imperative. After training, the cascade comprises multiple classifiers, each assigned to a specific stage of object detection.

The object detection process unfolds in stages. Initially, a cascade classifier is applied to determine whether an image contains the object of interest. If the classifier confirms the object's presence, the image progresses to the subsequent detection stage, where the next classifier comes into play. This iterative process continues until the final detection stage is reached. Each successive classifier operates on image regions identified by the preceding one, effectively excluding areas where the object was not detected. This intelligent cascade approach substantially reduces the number of regions requiring scrutiny at each detection stage, thus significantly accelerating the algorithm's performance.

When benchmarked against alternative methods like Eigenfaces, Fisherfaces, or LBPH [8], the Haar cascade classifier method emerges as the pinnacle of accuracy and efficiency in facial recognition. Eigenfaces and Fisherfaces tend to suffer from overfitting issues and exhibit relatively sluggish processing speeds, while Local Binary Patterns Histograms often falter in the face of variable lighting conditions, resulting in diminished recognition accuracy. Therefore, the choice of the Haar cascade classifier method for facial recognition and tracking within an Arduino-based system is a judicious one, particularly when confronted with limited computational resources.

The system used in the research was implemented and built on the Arduino Uno Rev3 microcontroller board based on the ATmega328P. Images were obtained from a Logitech C270 HD webcam operating in real-time mode, with a notable frame rate of 30 frames per second and a resolution of 1280 x 720 pixels, rendering sharp and fluid video captures. During the project, research was conducted to analyze the implementation of the system that was created for facial recognition and tracking. An analytical review of existing systems that had already been proposed on the market was conducted, and a general algorithm of their operation was created. In addition, a list of hardware and software components necessary for implementing the intended functionality was selected. The structural diagram of the device is shown in Fig. 2.

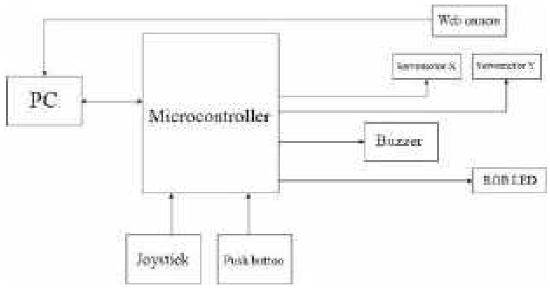


Fig. 2. The structural diagram of the device for face recognition and tracking

In order to effectively process images captured by the webcam, the system was meticulously designed utilizing the Python programming language in tandem with the computer vision library, OpenCV. The complementary hardware infrastructure of the system was brought to life through the utilization of an Arduino Uno board, complemented by the incorporation of corresponding modules. As depicted in Fig.

3, this ensemble of hardware and software elements coalesces to form the operational framework for our facial recognition and tracking system

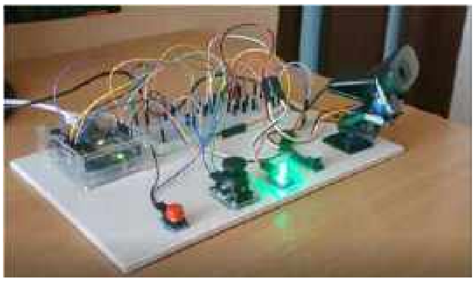


Fig. 3. The appearance of the system for face recognition and tracking

Delving into the algorithm that underpins this system, a detailed sequence of operations comes into focus. The process commences with the system's initialization, where it is imperative to reset the prior state of the servo motors, configuring them to a predefined standard position. This crucial preliminary step ensures that the servo motors embark on their task from a consistent starting point, thus establishing a level playing field for subsequent facial recognition endeavors. Only once this initialization process is complete does the system transition to the phase where it stands prepared to undertake facial recognition tasks. This meticulous attention to detail within the algorithm serves as the cornerstone of the system's reliability, guaranteeing that each recognition cycle begins on the same footing, fostering consistency and precision in the facial recognition and tracking process.

In a successful facial recognition endeavor, a defining feature is the visual feedback provided to the user: a bounding rectangle artfully encapsulating the detected face. This graphical confirmation is not merely a functional element; it is an interactive and user-friendly facet that instills confidence in the recognition process. Beyond its aesthetic appeal, this rectangle serves as the foundational region for subsequent analysis and processing of the identified face.

The critical pivot in this process is the adaptation of the servo motors' tilt angles based on the face's spatial position. This intricate stage entails the automatic and precise adjustment of tilt angles in such a manner that the system adeptly tracks the facial subject. Its core mission is to ensure that the face remains perfectly centered within the observation zone, regardless of the user's movements. This responsive action, initiated by the system, maintains the face within its field of view consistently.

As a testament to its performance, comprehensive testing yielded remarkable results. The facial detection accuracy consistently exceeded 80%, even in the presence of diverse challenges such as varying distances of up to one meter from the camera, fluctuating lighting conditions (ranging from adequate to insufficient), and potential tilting of the face in both horizontal and vertical orientations.

Moreover, the system's response time was found to be highly satisfactory, rendering it suitable for an array of facial recognition and tracking tasks. As depicted in Fig. 4, the chart illustrates the accuracy fluctuations of the facial recognition system across different distances from the camera, further elucidating the impact of lighting conditions. Evidently, accuracy dips with increasing distance and in subdued lighting. These trends can be attributed to the inherent challenges posed by reduced image quality at greater distances and the inherent difficulty in recognizing faces under suboptimal lighting conditions. Nevertheless, the system's overall performance remains commendable.

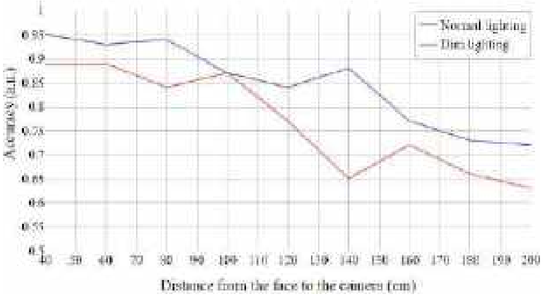


Fig. 4. The accuracy of the system as a function of distance under normal and subdued lighting conditions

Fig. 4 provides a comprehensive insight into the accuracy assessments of our facial recognition system, conducted under various conditions. These conditions encompass diverse distances from the camera and deliberate tilting of the faces both horizontally and vertically by 10 degrees. The data gleaned from these tests offers valuable insights into the system's performance under different circumstances. The discernible trend in the results is a decrease in system accuracy with an increase in the distance between the camera and the subject, as well as when the faces are tilted. These trends can be attributed to the pronounced shifts in perspective that occur as subjects move farther from the camera, resulting in alterations to the facial image's geometry. Similarly, tilting of the face introduces additional complexity, further challenging the system's recognition capabilities.

To provide a more granular view of these findings, Fig. 5 presents a graphical representation of the change in recognition accuracy with respect to distance, coupled with a 10-degree horizontal and vertical tilt of the face. This chart underscores the impact of distance and facial orientation on recognition accuracy, painting a vivid picture of the system's performance.

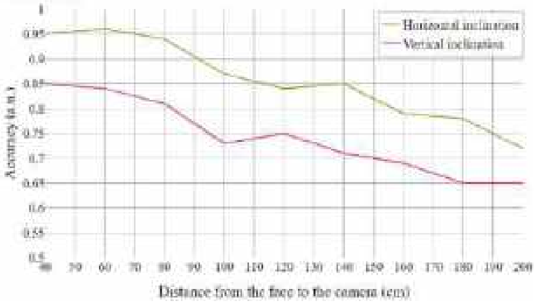


Fig. 5. The chart of system accuracy changes relative to the distance during a horizontal and vertical tilt of the face by 10 degrees

In general, the test results indicate that the accuracy of the facial recognition system depends on many factors, such as distance from the camera, lighting, and the angle of the face. These factors should be considered when designing and using a facial recognition system to achieve the highest accuracy. It is noteworthy that in both test cases, where lighting conditions were reduced and when the head was tilted by 10 degrees, the facial recognition system ceased to recognize the face at an approximate distance of 240-250 cm from the face to the camera. This critical boundary highlights the importance of maintaining optimal conditions within these parameters for the reliable performance of facial recognition systems.

Having embarked on this research journey centered around a facial recognition and tracking system based on the Arduino microcontroller, we can glean several insights that pave the way for potential improvements. It is imperative to acknowledge that crafting an optimal system for facial recognition and tracking poses a formidable challenge, demanding mastery over a spectrum of tasks encompassing real-time data processing and system efficiency. The pursuit of such improvements, however, remains an ongoing endeavor fueled by the quest for technological advancement and enhanced performance.

Firstly, the utilization of more robust microcontrollers with enhanced processing capabilities presents a promising avenue for improvement. Microcontrollers like the Arduino Mega or Raspberry Pi offer higher computational performance and power, enabling accelerated data processing and facial recognition tasks. Nevertheless, implementing such upgrades may face practical challenges. Integrating these more powerful microcontrollers might demand hardware modifications and software adjustments to harness their full potential.

Secondly, upgrading the system's imaging component by employing a more advanced webcam with increased resolution and data transfer speed can significantly enhance facial recognition accuracy. A high-resolution webcam provides finer-grained image details, enabling the system to capture subtle facial features. However, this enhancement comes with its own set of challenges. Ensuring seamless integration of the webcam with the existing system hardware and software can be complex and may demand custom drivers or software development.

Thirdly, to increase the system's speed, parallel data processing and asynchronous data transfer between the hardware and software parts of the system can be used. This reduces the time required for data processing and transmission between the parts of the system. With these improvements, more accurate and faster facial recognition can be provided, making the system more efficient and useful for use in various areas, such as security systems, video surveillance, advertising, and more. To ensure greater accuracy in facial recognition, neural networks and other machine learning algorithms can be used. For this purpose, it is necessary to have a large database of facial images used for training the neural network. After training the model, it can be used for real-time facial recognition.

To attain greater accuracy in facial recognition, incorporating deep learning techniques, particularly convolutional neural networks (CNNs), is a promising avenue. However, this approach poses its own challenges and limitations. Developing and training a neural network necessitates a substantial dataset of facial images for effective model learning, which can be time-consuming. Furthermore, exploring deep learning technology introduces the need for careful consideration of computational resource allocation. CNNs, while capable of delivering precise results, often require substantial processing power and memory, making them less suitable for real-time applications.

Overall, the facial recognition and tracking system based on Arduino microcontroller has great potential, but to ensure its greater efficiency and accuracy, more powerful microcontrollers, more productive programming environments, and optimization of data processing algorithms are necessary.

Conclusion

This scientific research project aims to investigate and evaluate the effectiveness of the Haar cascade classifier algorithm. To achieve this objective, a diverse sample of facial photographs was utilized, consisting of images captured at different distances from the camera, under different lighting conditions, and with different facial orientations within the camera's frame of view.

The experimental results demonstrated that the Haar cascade classifier method is highly accurate (more than 80% of successful attempts) in recognizing faces under various challenging conditions, such as diverse lighting and facial orientations. In addition, the algorithm proved to be highly efficient and demonstrated effectiveness in systems with limited computational capabilities. Unexpectedly, during the testing of the system, intriguing results emerged that worth further investigation. The algorithm displayed a resilience to variations in facial expressions, indicating potential robustness in real-world applications where individuals may demonstrate a wide range of emotional states.

Based on the findings of this research, it can be concluded that the Haar cascade classifier algorithm is a promising and effective solution for facial recognition systems. Moreover, the study uncovered various factors that can be considered in further improving the accuracy of the algorithm. To elaborate further, the Haar cascade classifier algorithm is a machine learning-based method that identifies facial features and patterns in digital images. It functions by analyzing a set of features, such as edges, lines, and corners, which are used to create a cascade of classifiers. These classifiers are then used to detect the presence of the facial features in an image.

References

[1] S. Singh Bhadauriya, S. Kushwaha, S. Meena, “Real-Time Face Detection and Face Recognition: Study of Approaches,” Lecture Notes in Networks and Systems. Singapore, 297–308, 2023.

[2] L. T. H. Phuc, H. Jeon, N. T. N. Truong, J. J. Hak, “Applying the Haar-cascade Algorithm for detecting safety equipment in safety management systems for multiple working environments,” Electronics, 8(10), 1079, 2019.

[3] M. Andrejevic, N. Selwyn, “Facial recognition technology in schools: Critical questions and concerns,” Learning, Media and Technology, 45(2), 115-128, 2020.

[4] X. Lai, P. L. P. Rau, “Has facial recognition technology been misused? A public perception model of facial recognition scenarios. Computers in Human Behavior,” 124, 106894, 2021.

[5] R. A. M. Budiman, B. Achmad, A. Arif, L. Zharif, “Localization of white blood cell images using Haar cascade classifiers,” 2016 1st International Conference on Biomedical Engineering (IBIOMED), 1-5, 2016.

[6] K. Berggren, P. Gregersson, “Camera focus controlled by face detection on GPU,” Department of Computer Science, Lund University, 2008.

[7] G. Sumanth, K. Kanimozhi, V. Murugesan, “Face Identity Detection and Recognition using Novel Convolutional Neural Network in Comparison with Haar Cascade to Improve Accuracy,” 14th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), 2022.

[8] I. U. W. Mulyono, A. Susanto, E. H. Rachmawanto, A. Fahmi, “Performance analysis of face recognition using eigenface approach,” 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), 1-5, 2019

Approaching Quantum Utility by Leveraging Quantum Software Stack

Markiian Tsymbalista
Faculty of Electronics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
markiian.tsymbalista@gmail.com

Mykola Maksymenko
Haiqu
Lviv, Ukraine
mykola.maksymenko@haiqu.ai

Ihor Katernyak
Faculty of Electronics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
ihor.katernyak@lnu.edu.ua

Abstract **— Improvement in Quantum Computing (QC) performance will allow us to solve a wide range of complex problems that classical computers of today can't handle. Nowadays we feel closer to achieving a state of Quantum Utility (QU) than ever before. Importance for both academia and business to understand the current state of technology and tooling, their extensibility points along with perspective optimization algorithms are on the verge. The purpose of the article is to provide a structured analysis of existing progress in QC. It serves a purpose of a guide on where significant progress is expected in the upcoming years, outlines details about tooling to start doing experiments, and introduces Quantum Computing Optimization Middleware (QCOM) reference architecture as a backbone around which new building blocks are expected to be added in the upcoming years, e.g., industry-specific enterprise connectors. The article reviews only the software stack, not considering opportunities for hardware as they seem to be much further away. This should help scientists and engineers to define a mental model of how to move forward to reach both mid and long-term goals toward QU.**

Keywords — quantum computing, quantum utility, quantum algorithm performance, Quantum Computing Optimization Middleware, error correction, qubit mapping.

I. Introduction

QC domain which includes hardware, software, tools, algorithms, etc. is usually perceived as more complex in comparison to classical technology. The knowledge base for the domain is limited and is mostly driven by research studies and technical documents of a few tools that were created by leaders of the industry. The next breakthrough in QC is going to happen at the intersection of algorithms (not high-level algorithms for solving specific problems from let's say physics domain, but algorithms on the compilation level of quantum computing stack) and existing tools (programming languages, software libraries or compilators). These programming languages and compilers have knowledge instilled of how to manipulate physical qubits on specific hardware, translate them to virtual ones, perform error correction and expose a high-level interface that hides most of the complexity to allow efficient implementation of QAs.

There is a number of publications around the basics of quantum computing and tools [1], publications that get under the hood of what is happening during compilation [2], studies of the performance of different compilers along with details of the tool that allow the combination of compilation steps from different vendors [3]. Also, studies that show optimization algorithms for circuit mapping [20], improvement of error correction [16], etc.

Despite all that, it is hard to make a conclusion on how to contribute to progress on promising directions and approaches to achieve QU, a term coined by in [29] to characterize practicality of modern quantum processing units (QPUs). Details of technological interfaces in tools around which performance optimizations could be tried are not summarized. The same for compilation layer phases with details of potential optimization benefits along with nascent algorithms of each phase that could contribute to breakthrough.

The purpose of the study is to analyze the current state of the industry and outline the most recent challenges and opportunities ascending in the field from the software perspective. This should help researchers and software developers to get a better feeling about potential ways of moving closer to what is called QU.

II. Methodology

Paper utilizes methods from qualitative research. Analysis of the most recent studies that have been conducted over the course of three years, to draw a picture of existing QC eco-system state and potential points of breakthrough. Also, 4 rounds of incremental interviews have been conducted with 3 thought leaders. Upon completion of every round, interview responses have been shown to all participants. Based on results from previous round response answers for the next have been refined. This approach helps to focus on the most important aspects and crystalize thinking about each of them. Central role in the process has been taken by co-founder of Haiqu [33] - Mykola Maksymenko [34] who helped to distil the course of thinking about the most recent challenges in QC, and provided guidance on tooling along with approaches that could be used to measure the performance of experiments in the space.

III. How to approach initial attempts of QC performance improvement. Extensibility of compilation pipelines.

In order to experiment with performance improvement of algorithms we need to get under the hood of compilation and customize different steps. As it is in every opinionated framework or tool, those choices are limited, but they significantly reduce overhead and allow to implement hypotheses quickly without a need to build a new compiler from scratch. From the other perspective it not going to work for more complex experiments. It is vital to provide a quick overview of some popular tools.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

A. *Qiskit Terra*

Transpiler provides extensibility points with its Transpiler Passes and Pass Manager [14]. The package allow us to write new transpiler passes (circuit transformations) and combine them with different existing passes, handling their order. The whole pipeline operates under the orchestrator (PassManager). It is responsible for communication between stages (passes) and their scheduling. It is worth mentioning various optimization levels provided through pre-defined pass managers that reflect additional complexity of compilation pipeline. There are optimization levels 0 to 3. The higher the number, the more optimized circuit will be generated. This number depends very much on actual hardware and of course on the algorithm.

There is a number of frameworks and tools that extend capabilities and quality of Qiskit compilation tools. For example, Fire Opal (one of the leading commercial product in quantum optimizations space) employs Qiskit's preset pass manager with optimization level 3 as its foundation and showcases consistent improvement of the result vs out-of-the-box optimization.

B. *Qiskit Pulse API [22].*

It is a pulse-level quantum SDK. This level allows more control when programming Quantum Hardware (QH). Achieving optimal QH performance necessitates instantaneous pulse-level instructions. Pulse delivers precisely that, empowering scientists to define experiment dynamics with precise timing. This SDK is particularly influential in enhancing error mitigation methodologies. A very insightful study of how Pulse API can be leveraged is provided in reference [19]. In this study, they utilized the Pulse API to practically evaluate numerically optimized error-robust pulses for implementing single-qubit gates on IBM's cloud-based quantum computers. The general concept of how API works is that arbitrary time-ordered signals (instructions) are provided as input, concurrently scheduled across multiple virtual hardware or simulator resources (channels). The system also facilitates the user in reconstructing the time dynamics of the measured output.

C. *Q# compiler extensions [23].*

For those, closer to the Microsoft eco-system, custom compilation steps allow us to extend and customize the Q# compilation process similar to the Qiskit Transpiler approach. Similar to any compiler, the Q# compiler follows a sequence of compilation stages. The initial steps involve parsing and validating the Q# code, and generating a data structure that captures the compilation's state. Subsequent stages traverse this data structure to generate updated versions. These stages often involve tasks like simple optimizations, constant folding, loop unrolling, as well as function and operation inlining. With custom compilation steps there is a possibility to get into the pipeline of execution, so custom, more complex optimizations could be done.

This is not an extensive list of the tools with plugin points, but they are sufficient to start trying custom optimization approaches.

IV. Quantum Computing Optimization Middleware

Most of the studies today look at the QC process from the perspective of layered architecture which has been inspired by the OSI networking model [10]. Since 2012 things have changed in the field and today reasoning about QC algorithm solely from this perspective doesn't provide detailed visibility of internals.

Q-CTRL pioneered the term "Quantum infrastructure software" [11] which they sell as a complex solution comparable to VMware, Citrix, and similar complex virtualization technologies in classical computing. These systems are heavier. The purposes of those systems are different for quantum and classical computing. Virtualization is a layer over the operating system in cases where a hypervisor is used. In QC this layer is not present, so middleware seems to be a better word to reason about the concept. QCOM (Fig. 1) is a better term to describe a software layer that is responsible for the performance optimization of arbitrary QAs. It shouldn't serve the purpose of doing commodity compilation/transpilation which is done via proxy QC Programming Framework (PF), but focus on enriching its capability, serving as a vitamin for researchers of QAs, to allow them to solve real-world problems with QC. It should be able to do optimization on both logical-level circuits and hardware circuits. The circuit processing in QC can basically be split into logical (higher-level theoretical computation) and physical (lower-level physical implementation) levels. A circuit at the logical level is defined using abstract, discrete-time gates and classical computations executed in real time. This is transformed by a compiler into a circuit at the physical level, where quantum operations are realized using continuous, time-varying signals adhering to specific timing constraints. There is still an uncertain question of whether proprietary compilers could allow for sufficient flexibility in the long run, so potentially QCOM could be responsible for compilation until both product segments will mature and have better, more standardized interfaces between each other.

QCOM could be exposed as a package plugin for different QC libraries, which calls a complex optimization layer

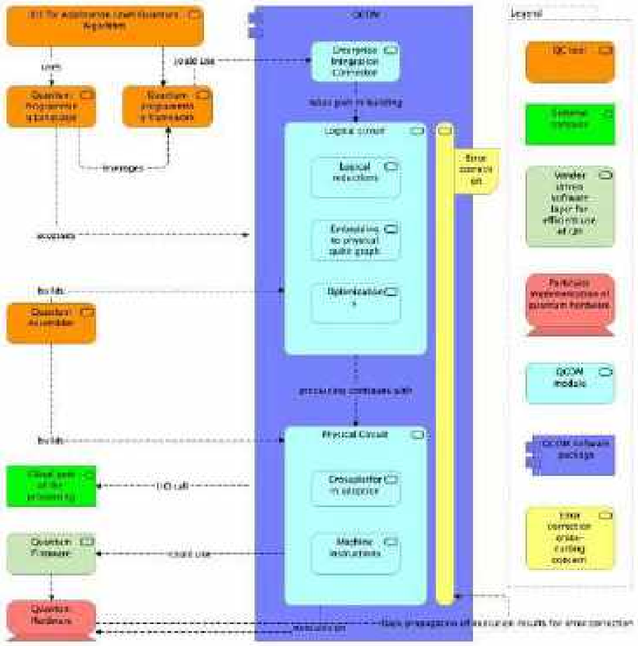


Fig. 1. Conceptual architecture of QCOM and its interactions with different QC tools, external components, and hardware.

running in the cloud. This software layer will be driving the industry in the upcoming decade. It will be called one way or another, but there will be dozens of companies working on it in some form, leveraging techniques from High-Performance Computing (HPC), Machine Learning (ML), and more classical QAs until the point of reaching QU where this component could grow into something else like let's say native to QH firmware.

A. Logical reductions.

Refer to techniques or processes used to simplify or transform a quantum circuit or logical expression into a more concise or manageable form while preserving its computational equivalence or logic. These reductions aim to make quantum computations more efficient, understandable, or amenable to analysis. They involve identifying patterns, redundancies, or equivalent operations within the quantum circuit or algorithm and simplifying them without changing the outcome. Logical reductions can encompass various techniques, such as gate fusion, gate commutation optimization, constant propagation, gate cancellation, simplification of logical expressions, etc.

B. Embedding to physical qubit graph.

Embedding to a physical qubit graph often referred to as qubit mapping or qubit allocation, is a crucial step in QC when running QAs on real quantum processors. The physical qubit graph represents the connectivity and layout of qubits on a specific quantum device. In quantum computers, qubits are not always fully connected to each other due to limitations in hardware design and qubit connectivity. This means that qubits can only directly interact with certain neighbouring qubits, and operations involving non-adjacent qubits need to be decomposed into sequences of single-qubit and two-qubit gates.

There are a lot of opportunities at the current level. For example, the paper [20] uses the circuit template matching optimization method which accounts for connectivity constraints of different topologies. Method satisfies those constraints by cutting the number of gates. The impressive thing about it is that it outperforms Qiskit across not only one but various IBM hardware architectures. It could serve as a baseline for further research in the direction that could incorporate more complex algorithms. Another family of approaches relies heavily on ML. Authors in the study [30] prove that it is possible for ML systems to learn based on historical data as there are many circuits executed on QH every day. These techniques are closer to cloud vendors which could instill that as part of their policy. Study [20] leverages a deep reinforcement learning approach that doesn't rely on batches of historical data but could learn along the way. Reinforcement learning techniques strive to acquire an action policy that dictates the appropriate action to take based on specific observations of the current state. The goal is to maximize a cumulative reward function. Within the proposed framework, observations are derived by extracting specific attributes from the quantum circuits at each state. Furthermore, a sparsely defined reward function is employed to indicate the achievement of a final state and subsequently assess the "quality" of the resulting circuit. This assessment could pertain to factors such as the resulting gate count, circuit depth, or anticipated fidelity.

Paper [9] concentrates on the problem of qubit routing leveraging a new decomposition approach based on the capabilities provided by integer programming, which also shows positive intermediate results. Qubit mapping focuses on the logical-to-physical qubit assignment, ensuring that the quantum circuit's logical qubits are placed on available physical qubits. Qubit routing, on the other hand, involves determining the pathways that qubit states will take as they move through the hardware during gate operations.

Research in that direction has been booming over the last couple of years with a lot of studies that use different approaches, so it is expected that it's going to have a near-term impact on reaching the state of QU.

C. Optimizations.

This is a layer that is responsible for the optimization and simplification of quantum circuits, particularly for near-term quantum devices with limited resources and high error rates. Some of the techniques that could be used:

- N-Qubit Blocks Clustering. This refers to the grouping or clustering of multiple adjacent or non-adjacent qubits that interact together in a quantum circuit. Clustering qubits that frequently interact in the circuit can reduce the need for repeated qubit swaps or additional gates, which can help mitigate errors and improve circuit performance. By optimizing the arrangement of qubits based on their interaction patterns, joint n-qubit blocks clustering aims to enhance the efficiency of quantum computations.
- 1-Qubit Optimization. This involves optimizing and simplifying the operations applied to individual qubits (1-qubit gates) within a quantum circuit. By minimizing the number of gates or finding gate sequences that are more robust against errors, 1-qubit optimization aims to improve the overall quality of the circuit. Effective 1-qubit optimization techniques can lead to more reliable quantum computations, especially on noisy devices.
- Blocks Consolidation. Refers to the process of identifying sequences of gates that can be combined or condensed into more efficient operations. By identifying patterns or sequences of gates that can be optimized, blocks consolidation reduces the overall gate count and complexity of the circuit. This can result in faster execution times and reduced susceptibility to errors.

These concepts collectively contribute to making quantum circuits more suitable for execution on near-term quantum hardware. They address challenges posed by limited qubit connectivity and noise. The goal is to create more compact, efficient circuits that can be executed with higher fidelity on currently available quantum processors.

D. Error correction cross-cutting concern.

In traditional software architecture, the cross-cutting concern is referred to a specific aspect of software that spans several logical layers of an application (logging, authentication, .etc). In QC error correction techniques are applied on every layer of program compilation and even as part of post-processing after QA execution, so it is a good candidate to be considered as one of the most important cross-cutting concerns of quantum software architecture.

Noise and errors in QH are one of the core challenges that arise due to the delicate nature of quantum systems and the influence of their surrounding environment. Considering that there are a lot of factors that cause errors it is worth summarizing what strategies could be considered to deal with them:

a) Error Suppression. Techniques aimed to prevent errors from happening. They reduce the likelihood of hardware error while quantum bits are being manipulated or used for memory storage. It leverages the nature of quantum control. More details could be found in a study [16]. Also, ML could help to increase the robustness of quantum gates [17]. Other strategies are described in [18] [19]. Most of the studies are done by the Q-CTRL team. The assumption is that they use those approaches as part of their commercial product FireOpal. This strategy is not effective for problems like "Energy Relaxation".

b) Error Mitigation. Errors could happen during algorithms execution and while measuring output. Various approaches have been devised to address them, aiming to enhance outcomes through postprocessing. These strategies encompass diverse methods such as randomized compiling, measurement-error mitigation, zero-noise extrapolation, and probabilistic error cancellation, yet they exhibit shared implementation principles. In general, error mitigation strategies entail running numerous slightly varied iterations of a target algorithm and subsequently combining outcomes. The adjustments made to the circuit can either be randomized or follow a predetermined algorithm. Some form of this is used in a leading product of Q-CTRL - FireOpal. But due to the high costs of running algorithms several times, for the time being, it is a less effective strategy in comparison to Error Suppression strategies.

c) Quantum Error Correction. This strategy entails the creation of algorithms that are aimed to identify and fix errors. In general, they work by implementing redundancy, distributing the state of qubits to other qubits. Then by checking helper qubits, it is identified whether an error happened or not. If yes, the correction could be applied. A huge disadvantage is the number of qubits that are required. As we all know, the number of qubits is very limited in today's hardware. So as of today, this is not a very effective technique until better approaches are found. This is a huge opportunity for "rockstar" researchers. More insights on the approach could be found in [15].

Thorough examinations have revealed that quantum error suppression currently offers the most compelling demonstrated advantages and optimal adaptability for integration with error mitigation and quantum error correction, culminating in outcomes that exceed the cumulative impact of individual components. Details on the success of their combination could be found in [18].

Also, recently in important milestone has happened [35]. The key discovery of research lies in achieving an error threshold of 0.8% for quantum error correction. This threshold is of paramount importance as it establishes the maximum error rate that a quantum system can withstand while still performing precise computations, as outlined in the research paper. This achievement places the protocol on par with the well-established surface code, which has held the highest error threshold for nearly two decades. The protocol has been developed around a family of Low-Density Parity-Check (LDPC) codes, renowned for their high encoding rate. It employs a comprehensive procedure for implementing fault-tolerant memory, which involves the measurement of syndrome cycles. These cycles require ancillary qubits and a specific circuit structure comprising nearest-neighbor controlled-NOT (CNOT) gates. From a technical perspective, the connectivity between qubits is organized in the form of a degree-6 graph, comprised of two edge-disjoint planar subgraphs. A practical demonstration illustrates that the protocol can maintain the integrity of 12 logical qubits across 10 million syndrome cycles, utilizing only 288 physical qubits with a 0.1% error rate. In contrast, achieving similar error suppression using the surface code would necessitate over 4000 physical qubits. The practical implications of this research are substantial, particularly for near-term quantum processors. The protocol's efficiency and compatibility with existing quantum hardware present a potential pathway to bridge the divide between current QC capabilities and the ultimate objective of fault-tolerant quantum memory.

E. Cross-platform adoption.

Layer that is responsible for building an abstraction over different hardware implementations. It is more of an enterprise, nice to have feature, so we are not covering it in the scope of this analysis.

F. Machine instructions.

The translation from quantum gates to pulse-level instructions is a complex and crucial step in making QAs executable on real hardware. It requires a deep understanding of the physical properties of the qubits, noise sources, and control mechanisms. Incorporates the following phases:

- Gates to Pulses (Gate Decomposition). The process begins by translating abstract quantum gates used in QAs into specific sequences of physical gates that are available on the QH. These physical gates are then further translated into corresponding time-varying signals (pulses) that control the behaviour of qubits during gate operations. Translating gates to pulses involves calibration and characterization processes to fine-tune the parameters of the pulses.

- Timing Resolution. Timing resolution involves determining the precise timing intervals for applying pulses to qubits during gate operations. High timing resolution ensures accurate execution of gates, reducing the likelihood of qubits losing their quantum states due to timing errors.

- Pulse Optimization. Pulse optimization focuses on finding the optimal pulse shapes, durations, and timings to achieve desired gate operations. Optimization techniques, often involving classical computations, are used to minimize errors and improve gate fidelity by shaping the pulses in ways that mitigate noise and imperfections in the hardware.

- Control Flow Optimization. Involves enhancing the sequence of quantum operations (quantum gates) within a quantum circuit to optimize its execution efficiency and mitigate errors. Quantum control flow refers to the ordering of quantum gates and the management of quantum states as they evolve through the circuit. Optimizing control flow in quantum computing aims to improve gate fidelity, reduce decoherence effects, and enhance the overall performance of quantum algorithms. Gate fusion, gate reordering, optimal control sequences, etc. are some of the examples of techniques used on this level.

This integrated process ensures that QAs are translated into executable instructions that account for physical hardware constraints, timing precision, pulse optimization, and control flow considerations. The goal is to maximize the reliability and efficiency of quantum computations on real quantum processors.

As part of this chapter comprehensive overview has been provided of steps that should be considered when building custom QCOM for both commercial and scientific purposes. In the next chapter, we will talk more about tooling, that allows us to implement those fine-grained manipulations.

V. Tooling to get into the aggressive scientific research of QC optimization. The backbone for building QCOM

Continuing the discussion of extensibility, none of the tools have good support for developing next-phase technology for QC. Despite some extensibility points, they don't offer enough control of QH and algorithms execution. Interaction only on the gates level abstraction doesn't provide the ability to manage lower-level control sequences with the purpose to do calibration, error mitigation, and characterization. Therefore, it becomes imperative to possess the capability to regulate the timing and establish links between quantum instructions and their corresponding pulse-level executions across different physical implementations or technologies used to create and manipulate qubits in QC. Timing features are especially important for the characterization of decoherence, crosstalk [24], dynamical decoupling [25], etc.

All those features are present in OpenQASM (Open Quantum Assembly Language) [26]. As a programming language, it is designed to serve the purpose of intermediate representation (IR). It is used by upper-tier compilers to interact with QH, enabling the depiction of an extensive array of quantum operations along with classical feed-forward flow control based on measurement results. It natively supports abstractions of logical and physical levels via specific semantics. Control flow instructions can be used to program repeat-until-success algorithms [27] and magic state distillation protocols [28]. There are also many other examples that show the potential of OpenQASM to get us closer to QU. Gate modifiers are another important mechanism that allows the creation of new gates based on existing ones. For example, modifier for inverting ads more readability which greatly raises optimization opportunities (it is hard to understand context when gates are decomposed).

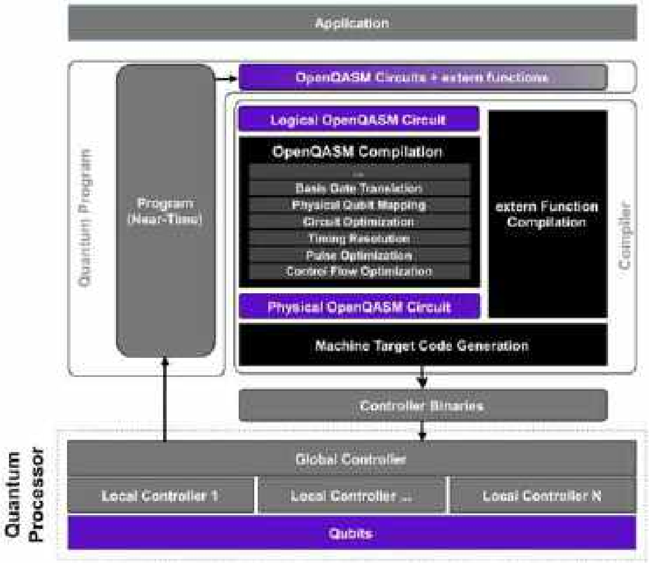


Fig. 2. The compilation and execution model of a quantum program, and OpenQASM's place in the flow [26].

A. *Program execution flow (Fig. 2).*

Specific segments are classical, amenable to writing in classical programming languages for near-time execution. The quantum program generates a payload for execution on QPU. This data package encompasses expanded quantum circuits and external real-time classical functions. External real-time classical functions refer to computational tasks or functions that are executed using classical computing resources, outside of the QPU. OpenQASM serves as the language for defining the quantum circuits, encompassing interface calls to external classical functions. There might be higher-order elements within the circuit data, subject to optimization prior to generating OpenQASM. An OpenQASM compiler can modify and optimize all aspects of circuits described through the IR, including basis gates, qubit mapping, timing, pulses, and control flow. The final physical circuit along with external functions is then processed by a target code generator to produce binaries intended for the QPU. Almost every aspect of program compilation could be controlled in detail. So as of today, OpenQASM can be considered a core tool for hard-core research of QC optimization and as a backbone for QCOM implementation.

Conclusion

This paper draws a picture of the current state of QC, looking into the most recent challenges and opportunities ascending in the field from the software perspective. QCOM conceptual architecture is proposed as an implementation backbone for optimization software spanning different layers of QA execution from trivial gate optimizations to more complex layout matching, manipulations on the machine-level instructions, and error correction. The motivation behind optimization on each layer is described along with interesting techniques that in combination increase the performance of the most popular tools like Qiskit. Analyses helped to identify priority directions for future research efforts, based on the recent progress in the industry, along with tools that could help to reach better control of quantum circuits execution. The proposed QCOM reference architecture aims to streamline the implementation of experiments and hypothesis testing in the field.

As the next steps in the research, prioritized directions that crystalized during this phase could be tackled: "Embedding to physical qubit graph" and "Error correction" (on both circuit and hardware layers). They showed a significant level of opportunity to reach QU in the nearest years. Covered optimization routines in different combinations could be implemented based on QCOM architecture, using OpenQASM programming language, and benchmarked using the Arline Benchmarks [32] tool to measure performance increase in comparison to well-established tools like Qiskit, etc. This will help to reason about how far we are from reaching QU and what could be done to get closer to it.

References

[1] H. Sahu, H. P. Gupta. "Quantum Computing Toolkit from Nuts and Bolts to Sack of Tools". 2023

[2] M. Maronese, L. Moro, Lorenzo Rocutto, Enrico Prati. "Quantum Compiling". 2021

[3] Y. Kharkov, A. Ivanova, E. Mikhantiev, A. Kotelnikov. "Arline Benchmarks. Automated Benchmarking Platform for Quantum Compilers". 2022

[4] Google Claims a Quantum Breakthrough That Could Change Computing. [Online]. Available:

https://www.nytimes.com/2019/10/23/technology/quantum-computing-google.html

[5] What is quantum supremacy? [Online]. Available:

https://www.techtarget.com/searchsecurity/definition/quantum-supremacy

[6] Youngseok Kim, Andrew Eddins, Sajant Anand, Ken Xuan Wei, Ewout van den Berg, Sami Rosenblatt, Hasan Nayfeh, Yantao Wu, Michael Zaletel, Kristan Temme, Abhinav Kandala. "Evidence for the utility of quantum computing before fault tolerance". 2023

[7] Why quantum 'utility' should replace quantum advantage. [Online]. Available:

https://techcrunch.com/2021/11/11/why-quantum-utility-should-replace-quantum-advantage/

[8] What Is NISQ Quantum Computing? [Online]. Available:

https://thequantuminsider.com/2023/03/13/what-is-nisq-quantum-computing/

[9] Friedrich Wagner, Andreas Bärmann, Frauke Liers, Markus Weissenbäck. „Improving Quantum Computation by Optimized Qubit Routing". 2023

[10] N. Cody Jones, Rodney Van Meter, Austin G. Fowler, Peter L. McMahon, Jungsang Kim, Thaddeus D. Ladd, Yoshihisa Yamamoto. "Layered architecture for quantum computing". 2012

[11] Quantum infrastructure software. [Online]. Available: https://q-ctrl.com/topics/quantum-infrastructure-software

[12] Differentiating quantum error correction, suppression, and mitigation. [Online]. Available: https://q-ctrl.com/topics/differentiating-quantum-error-correction-suppression-and-mitigation

[13] A. Paler, L. M. Sasu, A.-C. Florea, R. Andonie. "Machine learning optimization of quantum circuit layouts, ACM Transactions on Quantum Computing". 2022

[14] Transpiler Passes and Pass Manager. [Online]. Available: https://qiskit.org/documentation/tutorials/circuits_advanced/04_transpiler_passes_and_passmanager.html

[15] Quantum error correction. [Online]. Available: https://q-ctrl.com/topics/quantum-error-correction

[16] H. Ball, Michael J. Biercuk, Andre Carvalho, Jiayin Chen, Michael Hush, Leonardo A. De Castro, Li Li, Per J. Liebermann, Harry J. Slatyer, Claire Edmunds, Virginia Frey, Cornelius Hempel, Alistair Milne. "Software tools for quantum control: Improving quantum computer performance through noise and error suppression". 2020.

[17] Y. Baum, M. Amico, S. Howell, M. Hush, M. Liuzzi, P. Mundada, T. Merkh, A. R. R. Carvalho, Michael J. Biercuk. "Experimental Deep Reinforcement Learning for Error-Robust Gateset Design on a Superconducting Quantum Computer". 2021.

[18] P. S. Mundada, A. Barbosa, S. Maity, Y. Wang, T. M. Stace, Thomas Merkh, F. Nielson, A. R. R. Carvalho, M. Hush, M. J. Biercuk, Y. Baum. "Experimental benchmarking of an automated deterministic error suppression workflow for quantum algorithms". 2022

[19] A. R. R. Carvalho, Harrison Ball, Michael J. Biercuk, Michael R. Hush, Felix Thomsen. "Error-robust quantum logic optimization using a cloud quantum computer interface". 2020

[20] X. Gao, Zh. Guan, Sh. Feng, Y. Jiang. "Quantum Circuit Template Matching Optimization Method for Constrained Connectivity". 2023

[21] A.Zlokapa, A. Gheorghiu. "A deep learning model for noise prediction on near-term quantum devices". 2020

[22] Pulse. [Online]. Available:
https://qiskit.org/documentation/apidoc/pulse.html

[23] Extending the Q# Compiler. [Online]. Available:

https://devblogs.microsoft.com/qsharp/extending-the-q-compiler/

[24] J. M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, J. A. Smolin, J. M. Chow, C. A. Ryan, C. Rigetti, S. Poletto, T. A. Ohki, Mark B. Ketchen, M. Steffen. "Characterization of addressability by simultaneous randomized benchmarking. Physical Review Letters", 109, 24. 2012

[25] L. Viola, E. Knill, S. Lloyd. "Dynamical decoupling of open quantum systems". Physical Review Letters 82, 12. 1999

[26] OpenQASM Live Specification. [Online]. Available:
https://openqasm.com/intro.html#scope

[27] A. Paetznick, K. M. Svore. "Repeat-Until-Success: Non-deterministic decomposition of single-qubit unitaries. Quantum Information & Computation" 14, 15–16. 2014

[28] S. Bravyi, A. Kitaev. "Universal quantum computation with ideal Clifford gates and noisy ancillas". Physical Review A 71, 2. 2005

[29] N. Herrmann, D. Arya, F. Preis, Stefan Prestel. Quantum utility -- definition and assessment of a practical quantum advantage. 2023

[30] N. Quetschlich, L. Burgholzer, R. Wille. „Predicting Good Quantum Circuit Compilation Options". 2023

[31] MQT Predictor: Automatic Prediction of Good Compilation Paths. [Online]. Available: https://github.com/cda-tum/mqt-predictor

[32] Arline Benchmarks. [Online]. Available:
https://github.com/ArlineQ/arline_benchmarks

[33] Haiqu. [Online]. Available: https://www.haiqu.ai/

[34] M. Maksymenko LinkedIn profile. [Online]. Available: https://www.linkedin.com/in/mykola-maksymenko-4448a839/

[35] A. W. Cross, Sergey Bravyi, Jay M. Gambetta, Dmitri Maslov, Patrick Rall, Theodore J. Yoder . "High-threshold and low-overhead fault-tolerant quantum memory". 2023

Application of a Partitioned Nonlinear Controller for Crane Lifting Operations in Manufacturing

William R. Longhurst
Department of Physics, Engineering & Astronomy
Austin Peay State University
Clarksville, United States
longhurstw@apsu.edu

Randall D. Shelton
Department of Physics, Engineering & Astronomy
Austin Peay State University
Clarksville, United States

Bryan Gaither
Department of Physics, Engineering & Astronomy
Austin Peay State University
Clarksville, United States
gaitherb@apsu.edu

***Abstract* — Overhead cranes have historically been used in manufacturing and their operation has relied on human control. However, accidents due to human error have occurred. Automated control provides an opportunity for the reduction of accidents. The research presented in this article demonstrates the modeling and application of a partitioned nonlinear controller for the vertical lifting of loads. The control system is modeled around the actuator and the weight of the load. The partitioned aspect of the controller allows for optimum performance while lifting a wide range of load weights. Simulation and experimental results were obtained for step inputs and trajectory following. The results show critically damped to overdamped responses which were deemed desirable for safe operations. It was concluded the actuator-centered approach to control could be implemented on existing cranes in manufacturing. Particularly, it could be desirable for the automation of pick-and-place operations.**

Keywords — Material Handling, Lifting, Crane Operations, Manufacturing, Automation and Control

I. Introduction

Cranes have long been used in the manufacturing and construction industries to aid in the transfer of material or parts from one location to another. In manufacturing environments, the types of cranes that have historically been used are bridge, overhead or gantry style cranes. Their operation often entails the hoisting of a load up vertically and then moving it laterally to a desired spot where the load is then lowered and placed.

Because cranes hoist and lower loads with a cable, the load is susceptible to an undesirable swaying motion. In addition, the load could fall freely if tension is lost in the cable. These occur either through an external disturbance, unsmooth actuator motion, or collision with obstacles. The potential for load sway and loss of cable tension requires operator supervision and challenges the use of automation to control the crane. Historically cranes have been manually controlled by a human operator who can visually monitor and direct the movement of the load. Motion is directed by the operator through the use of two-position manual controls.

With the load prone to undesirable motion and the reliance on human supervision and control, accidents inevitably occur. As noted by Skiba [1], 90% of crane accidents in the United States can be attributed to human error. Additionally, Milazzo et al. [2] noted that poor human performance accounts for approximately 70-80% of the detected problems in crane operation.

The use of automation has the potential for improving safety and increasing the operational efficiency of cranes. The research presented in this article demonstrates the application of a partitioned nonlinear controller for the vertical lifting of loads. The work is novel because it presents an actuator-centered control method that can be implemented on existing cranes found in material handling systems. Unlike other approaches, the presented method allows for the weight of the load to be entered into the controller's algorithm without affecting the gain settings. Thus, optimum performance can be achieved for a wide range of load weights.

Previous experimental work has demonstrated the advantages of advanced control strategies when lifting and lowering loads vertically [3]. Prior knowledge of the load's weight or the ability to sense it at the start of a lifting operation was proven to be beneficial to a control system. However, the application of a linear controls limited the systems performance. Results from other works by Hermsdörfer et al. [4] as well as Polanen & Davere [5] have shown this to be true for humans as well as they lift objects.

Similar work by Celis et al. [6] studied the performance of a two-pulley-cable system that was used to lift loads vertically. The model for their study was based upon a design of an overhead crane. The study compared the system's responses from a state feedback controller and a proportional-integral-derivative (PID) controller commonly used in industry. Although more complex, the state feedback controller performed better by reducing oscillations. The authors concluded the PID controller generates unwanted behavior of the output when the system's dynamics are of the second order or greater.

Overhead, bridge and gantry style cranes use an overhead trolley to move in a 2-D plane motion while the load is lifted and lowered vertically by a cable. The vertically motion of the load is controlled independently of the trolley's motion. When transporting the cargo horizontally, the crane's cable will swing. Solutions involving feedback linearization by Tuan et al. [7] and Park et al. [8] have been presented to reduce the sway and improve the positioning of the cargo. Opportunity for further improvement to the control design was noted by Tuan et al. [7]. When the operation switched between lowering and lifting, the frequency response was different. In addition, Park et al. [8] with their controller design noted a weakness with handling model uncertainty and the importance of precise control of the cable's length.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

Cranes are used for lifting loads in a way similar to how other machines such as elevators are used. Elevators transport cargo vertically from one location to another and relay on a cable for the hoisting motion. Naturally there is a demand for improved operating efficiency of these vertical lifting systems. The demand for improvement can be met with advanced control systems and modeling. Similar to cranes, elevators are systems with varying load weights and motion control requirements. Lin et al. [9] showed the importance of understanding these variations. Similarly, work by Markon et al. [10] has shown the need for advanced and modern control strategies and how that can lead to optimum performance of elevators.

Advanced nonlinear modeling and experimental work for the control of overhead cranes has been presented with recent works [11], [12], [13], [14], [15]. Most of these works addressed the control of vertical and horizontal motion with a particular emphasis on the minimizing the load swing. Work by Bulín et al. [11] only addressed vertical motion but with attention to the dynamic interaction of the cable and the pulley. The work presents a modeling technique called absolute nodal coordinate formulation. Extensive work regarding nonlinear modeling and implementation of a controller was presented by Aguiar et al. [12] and Sorensen et al. [13]. Aguiar et al. presented a state-space fuzzy model for control and compared it to a quadratic controller. Their modeling and experimental results showed the state-space approach provided smoother motion. Sorensen et al. designed a controller that was comprised of three modules. The separate modules detected & reacted to positioning errors, detected and rejected disturbances and used input shaping to reduce oscillation of the load. When implemented on a large crane, the results showed a reduction of load swing. Work by Khatamianfar and Savkin [14] used state feedback control to independently control each joint of the overhead crane. Feedforward control was used in conjunction with the joint controllers to compensate for disturbances. With the use of real-time motion planning stable performance was observed. Sun and Fang [15] proposed a nonlinear control scheme that linearized the nonlinear dynamics. Detailed modeling and analysis were conducted to design the feedback controller. The control law was developed by modeling the mechanical energy of the system. Experimental results showed the load swing and disturbances were quickly suppressed.

Other recent works only included modeling and simulation in their study of cranes [16], [17], [18]. Asad et al. [16] proposed a fuzzy proportional–derivative (PD) based control strategy. The results were compared to a classical PD controller. Santhi et al. [17] used a quadratic regulator to control the transfer and swing motion of the load. The results were compared to a classical PD and PID controller which are commonly found in industry. All of the modeling was done using MATLAB - Simulink. The results presented by both Asad et al. and Santhi et al. showed improved performance over classical PD and PID control methods. Lastly, He et al. [18] designed a cooperative control method to control the load's position and regulate the swing or transverse cable deflection. With numerical simulations the authors were able to verify stable and robust control. When compared to a PD controller improved performance was demonstrated.

II. Modeling and Simulation

Inspiration for the partitioned approach to control was from work by Craig [19]. For this presented work, once the dynamic model of the physical crane system was developed, the control-law partitioning scheme of the controller was applied and gains selected to provide a critically damped response. A predicted response for the system was obtained through simulation using MATLAB – Simulink.

The physical crane system consisted a cable-pulley assembly and a DC motor that had a spool attached directly to its shaft as shown in Figure 1. As the spool was turned by the motor, the cable was retracted or fed through a system of pulleys. The cable-pulley assembly provided a mechanical advantaged and allowed for the lifting of heavy loads. The cable-pulley assembly consisted of an upper block of pulleys that was fixed in place and a lower block that was moveable and directly attached to the load. The mass of the pulleys as well as their rotational inertia were considered negligible and not included in the model. In addition, any friction acting in their bearings was considered negligible because of the relatively low angular velocities the pulleys experienced as the load was raised or lowered. Lastly, cable stretch was considered negligible as compared to the distance the load was raised or lowered.

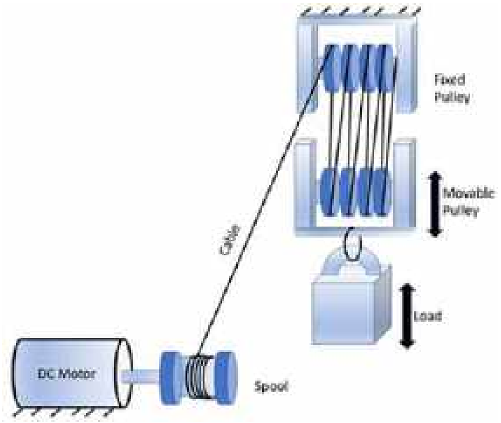


Fig. 1. A sketch of the modelled system

For our analysis, the coordinate systems and forces shown in the free-body diagram presented in Figure 2 were considered. Examining the forces acting on the load of mass m, one can see that it is subject to only its weight and the upward force F. The upward force F is provided by the cable-pulley assembly. Applying Newton's 2nd Law of Motion yields Equation 1. Solving for the force F yields Equation 2 which shows the force is a function of the load's acceleration $\ddot{x}$.

$$F - mg = m\ddot{x} \quad (1)$$

$$F = m\ddot{x} + mg \quad (2)$$

Considering the cable-pulley assembly shown in Figure 1 and the free-body diagram in Figure 2, the position x of the load can be related to the angular position θ_m of the motor's shaft and spool. The cable -pulley assembly has a mechanical advantage of n and the radius of the spool is R. The mechanical advantage of the cable-pulley assembly allows for the lifting of a larger load with a smaller input force applied to the cable by the spool & motor. This also means the spool & motor must move a greater amount of cable displacement through the pulleys than the load travels. In summary the mechanical advantage of the cable-pulley assembly enables the use of less applied force but requires a larger displacement of the cable by the spool. The relationship between the angular displacement of the motor & spool to the vertical position of the load is shown in Equation 3. To obtain the velocity and acceleration relationships, Equation 3 must be differentiated with respect to time. The resulting Equations 4 & 5 provide the velocity and acceleration of the load as a function of the motor's angular velocity and acceleration respectively.

$$x = \frac{R\,\theta_m}{n} \tag{3}$$

$$\dot{x} = \frac{R\,\dot{\theta}_m}{n} \tag{4}$$

$$\ddot{x} = \frac{R\,\ddot{\theta}_m}{n} \tag{5}$$

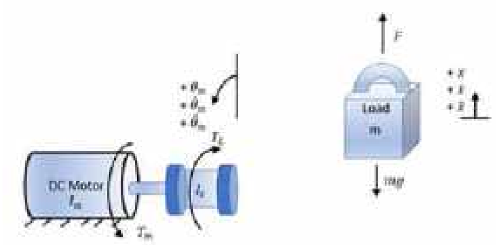


Fig. 2. A free-body diagram of the system

With the acceleration of the load given by Equation 5, it can then be substituted into Equation 2. This provides the lifting force from the cable-pulley system as a function of the motor's angular acceleration as well as the mass of the load.

$$F = m\left(\frac{R\,\ddot{\theta}_m}{n}\right) + mg \tag{6}$$

The load placed on the motor takes the form of torque. To determine the torque, we first need to determine the tension in the cable. This can be done by dividing the lifting force of the cable-pulley assembly by the mechanical advantage. The torque T_L caused by the load can then be calculated by taking the tension in the cable and multiplying it by the radius of the spool as given by Equations 7 & 8.

$$T_L = \left(\frac{F}{n}\right)R \tag{7}$$

$$T_L = \left(\frac{m\left(\frac{R\,\ddot{\theta}_m}{n}\right)+mg}{n}\right)R \tag{8}$$

Rearranging Equation 8 in the form of Eq 9 allows us to isolate the terms associated with the angular acceleration of the motor & spool. This will aid us later to build our controller model.

$$T_L = \frac{mR^2}{n^2}\ddot{\theta}_m + \frac{mgR}{n} \tag{9}$$

Referring to Figure 2, we can examine the free-body diagram of the motor & spool and develop an equation describing the motor's torque. From the Newton's 2nd Law of motion in angular terms the torque T_m provided by the motor is given by Equation 10. Equation 10 includes the inertia torques from the rotor and the spool as well as the torque from the load T_L. The mass moment of inertia I_s for the spool can be calculated using Equation 11 [20]. Further rearrangement and the substitution of Eq 9 into Equation 10 yields Equations 12, 13 & 14. Equation 14 provides the dynamic model of the physical system. From Equation 14 we can see the torque produced by the motor is nonlinear. The primary source of nonlinearity is the resulting weight of the load. This is evident with the last term of Equation 14. From Equation 14 it can be seen that the contribution to the motor's torque by the load is always unidirectional.

$$T_m = I_m\ddot{\theta}_m + b_m\dot{\theta}_m + I_s\ddot{\theta}_m + T_L \tag{10}$$

$$I_s = \frac{m_s}{8}\left({d_o}^2 + {d_i}^2\right) \tag{11}$$

$$T_m = (I_m + I_s)\ddot{\theta}_m + b_m\dot{\theta}_m + T_L \tag{12}$$

$$T_m = (I_m + I_s)\ddot{\theta}_m + b_m\dot{\theta}_m + \frac{mR^2}{n^2}\ddot{\theta}_m + mg\frac{R}{n} \tag{13}$$

$$T_m = \left(I_m + I_s + \frac{mR^2}{n^2}\right)\ddot{\theta}_m + b_m\dot{\theta}_m + mg\frac{R}{n} \tag{14}$$

As presented by Criag [19], we will partition the controller into two parts. The model-based portion will include all of the system parameters while the servo portion will only contain the controller gains and the error signals. For the model-based, the algorithm for control appears in the form of Equation 15. The torque from the motor T_m is set equal to a function contain the model parameters α & β and input from the servo T'. The goal for the model-based portion is to cancel the nonlinearities of the system. Thus, Eq 15 must be identical to the dynamic model presented in Equation 14. This requires the model parameters to be established according to Equations 16 & 17.

$$T_m = \alpha T' + \beta \tag{15}$$

$$\alpha = I_m + I_s + \frac{mR^2}{n^2} \tag{16}$$

$$\beta = b_m\dot{\theta}_m + mg\frac{R}{n} \tag{17}$$

With the establishment of the model parameters, the input from the servo T'must be set equal to the angular acceleration $\ddot{\theta}_m$ of the motor & spool as given in Equation 18. In its stated form, Equation 18 is the angular equation of motion for a system with a unit mass moment of inertia. In essence, the model-based portion of the controller has the effect of making the system appear to only have a unit mass moment of inertia.

$$T' = \ddot{\theta}_m \tag{18}$$

Since the we are seeking closed-loop control of the system, we will use the feedback of the angular acceleration, velocity and position and comparison to the desired values to establish the error signals e, $\dot{e}$, & $\ddot{e}$. These are given in Equations 19, 20 & 21 and will be used to establish the servo control law.

$$\theta_d - \theta_m = e \tag{19}$$

$$\dot{\theta}_d - \dot{\theta}_m = \dot{e} \tag{20}$$

$$\ddot{\theta}_d - \ddot{\theta}_m = \ddot{e} \tag{21}$$

With the appearance of the system to have a unit mass moment of inertia, the servo portion will be designed to control it. Equation 22 presents the servo control law. The input into the model T' is set equal to a function that contains the errors signals e & $\dot{e}$ as well as their respective control gains K_p& K_v. The additional term of the desired angular acceleration $\ddot{\theta}_d$ is included. Equating Equations 18 & 22 and then using the angular acceleration term instead of the input signal produces Equation 23 which then can be written in the form of Equation 24 with the use of Equation 21. The differential equation shown in Equation 24 determines the closed-loop response of the system. With the model-based portion all other terms have been cancelled.

$$T' = \ddot{\theta}_d + K_v\dot{e} + K_p e \qquad (22)$$

$$\ddot{\theta}_m = \ddot{\theta}_d + K_v\dot{e} + K_p e \qquad (23)$$

$$\ddot{e} + K_v\dot{e} + k_p e = 0 \qquad (24)$$

The block diagram of Figure 3 illustrates the system model and controller. From the block diagram, one can see the servo portion resides outside of the model-based portion. With the appropriate setting of the control gains, any desired response can be obtained.

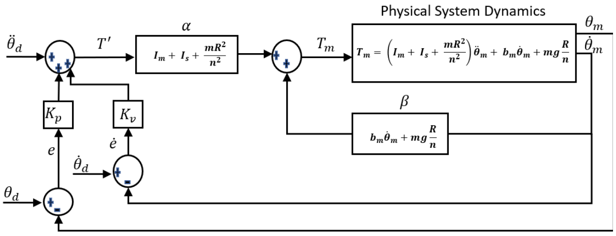

Physical System Dynamics
α
β
$\ddot{\theta}_d$
T'
$I_m + I_s + \frac{mR^2}{n^2}$
T_m
$T_m = \left(I_m + I_s + \frac{mR^2}{n^2}\right)\ddot{\theta}_m + b_m\dot{\theta}_m + mg\frac{R}{n}$
θ_m
$\dot{\theta}_m$
$b_m\dot{\theta}_m + mg\frac{R}{n}$
K_p
K_v
e
$\dot{e}$
$\dot{\theta}_d$
θ_d
+
-

Fig. 3. A block diagram of the system

The predicted response of the system to a step input is shown in Figure 4. The simulated system responded to a commanded signal to raise a 0.425 kg load 5 cm and then lower the load back to its original position.

As can be viewed in Figure 4, the modeled system had a critically damped response and was uniform for the upward and downward motions. This indicated the model-based portion of the controller cancelled the nonlinearity and performed as desired. The predicted response time was approximately 1 second but this could be changed with a different set of gain values.

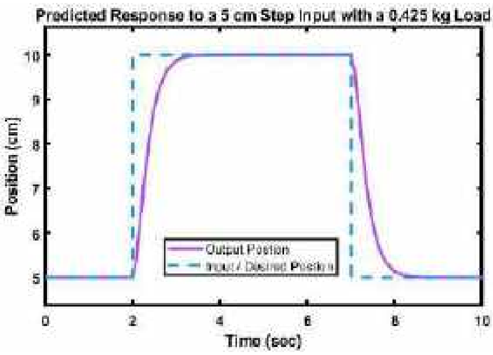


Fig. 4. The predicted response from the simulation

III. Experimental Setup

The crane used for the experimental portion of the research is shown in Figures 5 & 6. It consisted of a box frame with dimensions of approximately 0.67 m deep x 0.72 m wide and 1.86 m in height. The frame was constructed using T-slot aluminum extrusions and the corresponding connecting hardware. All of the crane's mechanical and electrical components necessary for automation of the lifting process were mounted onto the frame with the exception of the computer used for control and data logging. The implementation of the controller was accomplished using MATLAB – Simulink.

Lifting was accomplished by using a cable-pulley system shown in Figure 5. The cable-pulley system consisted of two light weight pulley blocks and cable. The upper pulley block was attached to the aluminum frame while the lower block was configured to allow for vertical movement. A light weight cable was selected that had little very little stretching capability. The cable was threaded through the upper and lower pulley blocks in a manner that provided a mechanical advantage factor of 8. The load consisted of known masses and were attached to the lower pulley block. For the experimental work three different loads were used: 0.275 kg, 0.425 kg and 0.575 kg.

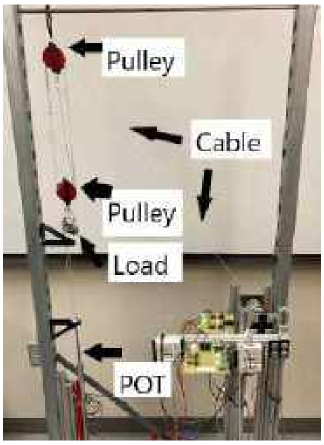


Fig. 5. View of the cable-pulley system

In addition to the load, an Omega LD621 displacement transducer with a 0.3 m stroke was attached to the lower pulley block. The displacement transducer functioned similar to a potentiometer (POT) and produced an 0-10 V signal that was proportional to the position of the lower pulley and load.

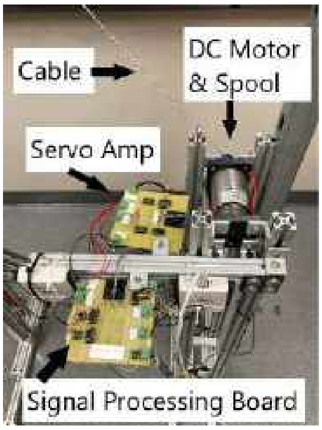


Fig. 6. View of the DC motor, servo amp, and the signal processing board

The actuator used for the crane was a Pittman Express 12 V brushed DC motor (model number 14203S010). It was end-mounted to the aluminum frame as shown in Figure 6. Attached to the motor's output shaft was a shaft coupling. The shaft coupling was modified to function as a spool for the cable.

Actuation of the DC motor was accomplished with a servo amplifier. With the use of the servo amplifier, a command signal of +/- 5V was used produce a desired motor torque T_m. The electrical components consisted of Fairchild TIP 125 PNP & TIP 120 NPN Darlington Transistors, a Texas Instruments LF-412 Operational Amplifier and three resistors.

With an input voltage of V_{in} the servo amplifier created a load current I_L through the DC motor. The relationship between the motor torque and the current passing through it is given by Equation 25 [19]. The motor torque T_m is directly proportional to the load current I_L. The proportional constant k is the motor's torque constant. For the DC motor used in the experimental setup, the torque constant k was 3.27 x 10^{-2} N-M/A. This was provided by the motor manufacturer.

$$T_m = kI_L \tag{25}$$

IV. Results

For comparison to the simulation model, the system's response to a 5 cm step input was evaluated. Upon initial testing a steady state error was observed. The amount of error was approximately 20%. This was based upon the crane being able to lift the 0.425 kg load a vertical distance of approximately 4 cm when prompted with a 5 cm step input. When lowering the same load, a similar error was observed. The source of the error was most likely unmodelled friction and mass of the cable pulley system.

To correct the error an integral component was added to the control law of Equation 24. The integral component continuously summed the amount of error present at each update cycle of the controller and multiplied it by a gain factor. However, to enable stability, two separate gain values were used. One gain was used when a positive error signal was present and the other when a negative error signal was present. The positive error signal corresponded to raising the load while the negative error signal corresponded with lowering the load. For the operation of raising the load, the value of gain selected was 0.25. For lowering the load, the value of gain was 0.15. These gain values were selected using empirical testing. These two values of gain were intentionally chosen in order to move the load as quickly as possible but insure safe operation.

Figure 7 presents the system's response to 5 cm step inputs. The step inputs were comprised of two raising and two lowering operations. As can be seen in Figure 7, the system did achieve the desired output position each time. The response can be described as critically damped but trending toward overdamped. It clearly was not underdamped. The observed response would be desirable for automated cranes that conduct pick-and-place operations. When compared to the response time in Figure 4, which presents the simulation's predicted response, the response time in the constructed crane system was longer. As can be seen in Figure 4, the predicted response time was approximately 1 second to lift or lower the load 5 cm. From Figure 7, the observed response time for the initial lifting operation was approximately 5 seconds while the response time to the second lifting and all lowering response times was approximately 10 seconds. It can also be observed in Figure 7 during the initial lifting operation, the system achieved approximately 80% (20% error) of the desired output in approximately the first second of lifting. At that point the rate of change of the position with respect to time slowed. This was the point where the system began to rely on the integral component in the control law to correct the remaining error present. As the value of the summed error continuously grew at each update cycle of the controller, the actuator responded with more motion until the desired output was achieved. As discussed prior, the slower motion near the desired output position was desirable for safe and stable operation. At the beginning of the initial lifting operation in Figure 7, the value of the summed error in the integral component was zero. Once the system achieved its initial step position of 10 cm, the summed error in the integral component of the control law had a positive value. The summed error remained present in the control law even though the current value of error was zero. Upon the commanded step to lower the load to the 5 cm position, the summed error signal was initially positive due to the prior upward motion. It took approximately 10 seconds for the summed error in the control law to grow to the required negative value that induced the additional motion in the actuator to completely overcome the remaining error present. This was true for all motions after the achievement of the initial step in position. When the direction of motion was reversed there was a longer response time to overcome the error present. This was due to the summation of both positive and negative error that was fed into the integral component of the control law. The longer response time as compared to the simulation model was a consequence of the integral component in the control law. However, for crane operations, the resulting slower movement can be beneficial and a key enabler for safe operations.

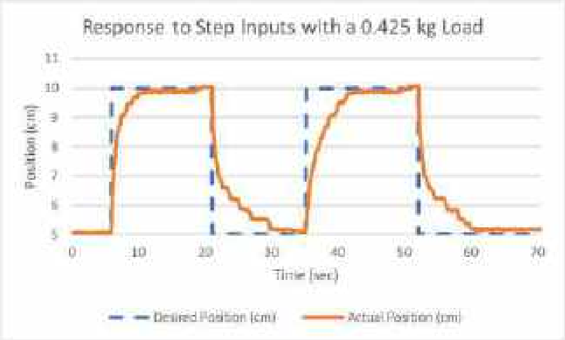


Fig. 7. Response to step inputs

Figure 8 presents the system's response to following a manually generated trajectory. As can be observed in Figure 8, the system was able to track the trajectory. However, adequate time was needed for the controller with its integral components in the control law to overcome error. The initial movement upward was tracked very closely with a lag of approximately 1 second or less. However, once the direction of the generated trajectory was reversed, a larger lag in the response emerged. If the trajectory underwent two reversals in direction in a relatively short time period, there was a significant deterioration in the tracking accuracy. This is evident in the 20 – 30 second time region of Figure 8. The desired position of the load was lowered from 15 cm to 8 cm in approximately 5 seconds. From the 25 second to 30 second region of time the desired position remained at 8 cm. The system responded by lowering the load but there was a delay of approximately 1-2 seconds in the response. The delay can most likely be attributed to the summed error in the integral component and the time needed to overcome the error. Before the system had positioned the load at the desired 8 cm, the desired position was raised to 13 cm beginning at the 30 second time point. At this point there was approximately 1 cm of position error. However, given enough time, the controller would move the load to the updated position in the desired trajectory as evident at the 40 second and 55 second time points in Figure 8. The tracking of a trajectory was very good as long as the direction was not reversed. Once reversed, the trajectory generated must not change too rapidly for a reasonable amount of accuracy to be maintained.

Similar results were obtained when the load was changed to 0.275 kg and 0.575 kg. Thus, indicating the system's ability to respond almost identically regardless of the weight of the load.

From the experimental setup and results presented, it can be concluded that the actuator centered control method provided an acceptable level of performance for crane operations. In addition, and because it is actuator centered, it could be implemented on existing crane systems found in manufacturing and construction. Particularly, it could be desirable for the automation of pick-and-place operations.

Applying the partitioned nonlinear controller allowed for the nonlinear elements in the system to be canceled by the model-based portion of the controller. The critical damping response was dictated by the servo portion of the controller. By partitioning the controller, different load weights can be lifting while still achieving the same critically damped response.

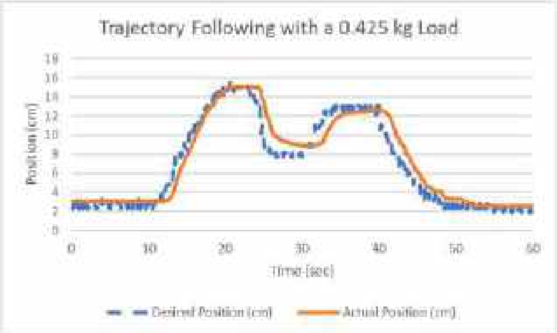


Fig. 8. Response to a manually generated trajectory

Because the approach to control was actuator centered, some friction and inertia elements within the physical system went unmodeled in the controller. To address this, an integral component was added to the control law that resides in the servo portion of the controller. Although the integral component eliminated the error, the system responded slower. However, slow movement of the suspended load is often desirable in crane operations.

References

[1] R. Skiba, Best Practice Standards and Methodology for Crane Operator Training – A Global Perspective, Journal of Transportation Technologies, Volume 10, Number 3, pages 265-279, July 2020.

[2] M. Milazzo, G. Ancione, V. Brkic, , Safety in Crane Operations: An Overview on Crane – Related Accidents, 6th International Symposium on Industrial Engineering, Belgrade Serbia, September 24-25, 2015.

[3] W. Longhurst, K. Prue, B. Gaither, Experimental Investigation into the Basic Application of Force and Position Control for Human-Machine Team Lifting Operations in Manufacturing, Proceedings of the Institution of Mechanical Engineers, Part B, Journal of Engineering Manufacturing, Volume 236, Issue 3, pages 174 - 189, February 1, 2022.

[4] J. Hermsdörfer, Y. Li, J. Randerath, G. Goldenberg, S. Eidenmüller, Anticipatory Scaling of Grip Forces When Lifting Objects of Everyday Life, Experimental Brain Research, Vol 212, pages 19-31, 2011. DOI:10.1007/s00221-011-2695-y.

[5] V. Polanen, M. Davare, Sensorimotor Memory Biases Weight Perception During Object Lifting, Frontiers in Human Neuroscience, December 23, 2015. DOI: 10.3389/fnhum.2015.00700.

[6] C. Celis, D. Amaya, O. Ramos, Design and Control of a System for Lifting Loads, Using State Feedback and PID Controllers, International Journal of Applied Engineering Research, Volume 13, Number 2, pages 1001-1006, 2018.

[7] L. Tuan, A. Janchiv, G-H. Kim, S-G. Lee, Feedback Linearization Control of Overhead Cranes with Varying Cable Length, International Journal of Precision Engineering and Manufacturing, 11th International Conference on Control, Automation and Systems, IEEE, Gyeonggi-do, Korea, October 26-29, 2011.

[8] H. Park, C. Dongkyoung, K-S. Hong, A Feedback Linearization Control of Container Cranes: Varing Rope Length, International Journal of Control, Automation, and Systems, Volume 5, Number 4, pages 379-387, 2007.

[9] K. Lin, S. Lupin, Y. Vagapov, Analysis of Lift Control Systems Strategies Under Uneven Flow of Passengers, International Federation for Information Processing - Advances in Information and Communication Technology, volume 470, pages 217-225, 2016.

[10] S. Markon, K. Aoki, M. Nakagawa, T. Sudo, Recent Trends in Elevator Group Control Systems, 23rd International Technical Conference on Circuits / Systems, Computers and Communications, pages 697-700, Yamaguchi, Japan, July 6-9, 2008.

[11] R. Bulín, M. Hajžman, P. Polach, Nonlinear Dynamics of a Cable-Pulley System Using the Absolute Nodal Coordinate Formulation, Mechanics Research Communications, Volume 82, pages 21-28, 2017.

[12] C. Aguiar, Leite, D., Pereira, D., Andonovski, G., Škrjanc, I., Nonlinear Modeling and Robust LMI Fuzzy Control of Overhead Crane Systems, Journal of the Franklin Institute, Volume 358, pages 1376- 1402, 2021.

[13] K. Sorensen, W. Singhose, S. Dickerson, A Controller Enabling Precise Positioning and Sway Reduction in Bridge and Gantry Cranes, Control Engineering Practice, Volume 15, pages 825-837, 2007.

[14] A. Khatamianfar, A. Savkin, Real-Time Robust and Optimized Control of a 3D Overhead Crane System, Sensors, Volume 19, Number 15: 3429, 2019. https://doi.org/10.3390/s19153429

[15] Sun, N., Fang, Y., Nonlinear Tracking Control of Underactuated Cranes with Load Transferring and Lowering: Theory and Experimentation, Automatica, Volume 50, pages 2350 – 2357, 2014.

[16] S. Asad, M. Salahat, M. Zalata, M. Alia, A. Rawashdeh, Design of Fuzzy PD-Controlled Overhead Crane System with Anti-Swing Compensation, Engineering, Volume 3, Number 7 pages 755-762, 2011.

[17] L. Santhi, M. L. Beebi, Position Control and Anti-Swing Control of Overhead Crane Using Optimal Control, International Journal of Industrial Electrical and Electrical Engineering, Volume 3, Issue 11, pages 28 – 33, 2015.

[18] W. He, S. Ge, Cooperative Control of a Nonuniform Gantry Crane with Constrained Tension, Automatica, Volume 66, pages 146 – 154, 2016.

[19] J. Craid, Introduction to Robotics Mechanics and Control 3rd Edition, Pearson Prentice Hall, 2005.

[20] R. Hibbeler, Engineering Mechanics – Dynmaics 13th Edition, Pearson Prentice Hall, 2013.

Pitch and Heading Angles Estimation by Airplane Trajectory Data Available in ADS-B

Ivan Ostroumov
Air Navigation Systems Department
National Aviation University
Kyiv, Ukraine
0000-0003-2510-931

***Abstract* —A number of Automatic Dependent Surveillance-Broadcast (ADS-B) applications is increasing every year. Nowadays each airspace user should be equipped with a transponder of mode 1090ES to identify itself in the air traffic. Vary of settings of on-board transponders provides a different list of parameters for sharing. The most common parameters include a unique identification code of user and a position report including latitude, longitude, and pressure altitude. Missing airplane orientation angles in ADS-B data set could be a problem for tasks of air traffic surveillance and airplane visualization in flight simulator software. Also, orientation angles could improve performance of trajectory filtering and prediction of airplane position in real-time data processing. In the paper, we consider calculation of pitch and heading angles based on trajectory data obtained by ADS-B. Also, analytical formulas for airplane orientation angles calculation based on ADS-B data are proposed in the paper. ADS-B trajectory data usually are not synchronized and include multiple gaps. A linear regression model with B-splines functions is used for trajectory data interpolation for a given time series. Also, surveillance data are used for vertical and ground speed calculation. Calculated orientation angles and interpolated position data form an array of six Degrees of Freedom (6DoF), which is widely used in different visualization tools.**

Keywords — trajectory data processing, surveillance, ADS-B, Array 6DoF, civil aviation, airplane, orientation angles

I. Introduction

Surveillance is an important component of modern civil aviation system. Surveillance is provided with the help of different on-board and ground-based systems and sensors [1, 2]. Services of air traffic control widely use ground primary and secondary radars to measure location of each airspace user [2, 3]. Automatic Dependent Surveillance-Broadcast (ADS-B) is a modern surveillance technology that is considered as a main for various applications in the future air navigation systems. According to ADS-B any airspace user have to be equipped with specific equipment to transmit identification signal omnidirectionally. Anyone within a range of wireless communication link can receive this data and use it for its surveillance functionality. Shared by ADS-B data includes coordinates of airspace user position which is measured by on-board navigation sensor and identification code of airspace user [4]. Each message of ADS-B is transmitted in open data format, therefore anyone could receive and decode it easily. Low-cost Software-Defined Radio (SDR) could be used to receive and collect air traffic data in real-time from airspace users located in the range of operation [5, 6]. Increasing role of ADS-B surveillance technology in future air navigation systems is a subject of cybersecurity analysis.

Different degradation factors may act in ADS-B data [7, 8]. Some of them could reduce performance of on-board positioning system, another can cause blocking digital messages data transferring channels [9, 10]. As an example interference with other digital messages issued at the same communication channel. Modified transponder of mode 1090ES (Extended Squitter) is the most frequently used onboard equipment of ADS-B. This transponder uses a 1090 MHz channel for data transmission. The 1090 MHz radio frequency is also used by secondary surveillance radars for range measuring and by Traffic Collison and Avoidance systems for addressed communication. The capacity of this data channel is highly limited in congested airspace with multiple users [11]. Thus, trajectory data obtained by ADS-B is not synchronized in time, which reduces the performance of surveillance data processing algorithms that are used in air traffic control. The accuracy of airplane coordinates obtained by ADS-B is associated with the performance of on-board positioning sensor. A receiver of the Global Positioning System is used as a primary navigation system on-board. However, in case of some faults, an inertial navigation system [12] or methods of positioning by navigational aids could be used [13, 14] based on criteria of maximum performance in some areas of airspace and meeting requirements of valid navigation specifications developed by air navigation services provider [15].

Another important problem connected with surveillance data is missed airplane orientation in air space. Euler angles of pitch, roll, and yaw compose a vector of airplane orientation in 3D space. Airplane position in latitude, longitude, and altitude together with angles of pitch, roll, and yaw form an array of six Degrees of Freedom (6DoF) [16]. Array 6DoF is widely used in different filters at the level of trajectory data processing to reduce level of noise in coordinates. Also, Array 6DoF is required for precise simulation of air traffic and for automatic flight phase identification [17]. Array 6DoF is used in the airplane model to predict delays of each airspace user at a particular waypoint of preplanned trajectory.

Array 6DoF is also required for airspace user flight visualization based on ADS-B input in different air traffic simulation tools. FlightGear, X-Plane, and other flight simulators could be used for visualization of 3D model of airplane trajectory for solving a variety of safety tasks [18], [19]. In this case, a full array 6DoF including complete series of data could be used [20]. ADS-B trajectory data includes readings at different times with changed periods. Therefore any ADS-B trajectory data should be synchronized in time by some math method of data interpolation. Then synchronized trajectory data could be used for orientation angles estimation and forming array 6DoF.

Orientation angles are measured on-board by Attitude Heading and Reference System and can not be read by any

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

ground sensor. In the paper, we analyze a possible way for angles of pitch, and yaw estimation based on data transferred by ADS-B. Also, we use a linear regression model with spline functions to interpolate missing data to a defined time series with a constant period.

II. Trajectory Data Processing

According to ADS-B, airplane on-board transponder of Mode 1090ES transmits digital messages via an omnidirectional antenna system. These signals could be received by ground networks of SDRs, by a constellation of Low Earth Orbit satellite of ADS-B receivers, or by on-board equipment of ADS-B IN [21]. Communication data channel uses messages of 112 bits long. There are three main types of ADS-B messages [22, 23]:

- Airspace user identification (includes identification call sign of airspace user and category of airplane used);
- Surface or airborne position (includes latitude, longitude, barometric altitude, or GPS height);
- Velocity (include vertical rate, ground, and airspeed);
- Operational status and navigation categories of accuracy.

According to the setting of on-board transmitter, different types of messages could be generated with different repetition frequencies. Each digital message includes a unique airspace user code of registration with the International Civil Aviation Organization (ICAO). The validity of successful data transmission is controlled by a cyclic redundancy check (CRC) [24].

Most parameters in ADS-B are transmitted in codded form to reduce the total length of data. After receiving a message of ADS-B with SDR it is decoded and saved in a database with a precise timestamp.

In the common case database of received messages include ICAO identification code, latitude, longitude, barometric altitude, and timestamp of fixed data. A full airplane trajectory may include long gaps due to airplane placement out of the maximum range of wireless communication links. Many areas around the globe are still out of SDR coverage. Also, many messages could be broken during data transfer. As an example, a few trajectories of flights are presented in Fig.1 based on ADS-B trajectory data. Detailed information about these flights is presented in Table I.

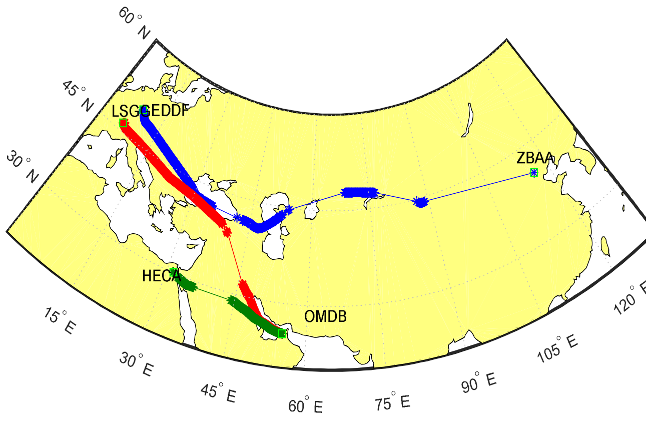


Fig. 1. Common ADS-B trajectory data.

TABLE I. Performance of ADS-B Trajectory Data Sets

Flight number	Departure/ Destination airports	Airplane type	Date	Flight time hh:mm	Dataset length
UAE83	OMDB/ LSGG	B 777-300ER	15-May-2023	06:52	611
DLH720	EDDF/ ZBAA	A340-300	15-May-2023	10:53	497
UAE927	OMDB /HECA	A380-800	26-May-2023	04:09	234
DLH760	EDDF/ VIDP	B747-400	2-June-2023	07:25	374
DLH489	EDDF/ MMX	B747-8	2-June-2023	11:46	1027
THA921	EDDF/ VTBS	B777-300ER	2-June-2023	10:46	734

Results in Table I show that poor ground network configuration (for these examples mostly in Asia) of SDR caused losing about 60% of data. For the case of THA921 on-board transponder of ADS-B generates one message per 20 s, which should give 1908 data points for the whole trajectory. However, only 734 messages are available for data processing which is 38.5%. The full data series should include 38760 measurements for each second of flight. Thus, before orientation angles calculation, it is strongly required to perform data recovery or interpolation for the required time series.

There are two basic steps of airplane trajectory data processing. In the first one, trajectory data should be recovered for synchronous time series. A polynomial function of different orders or spline functions could be used for data interpolation for particular time series. Also, different tracking filters like αβγ filter or Kalman filter could be used to minimize errors in on-board positioning system [25, 26].

It is also important to note that trajectory data obtained by ADS-B is accompanied by barometrical altitude and geometrical height is usually missed. Barometric altitude is calculated on-board by results of static pressure measurements. This altitude is counted from the standard pressure isobaric line. During the airplane placed in the airport vicinity, a digital weather report could be used to get transformation of barometric altitude to GPS height. For other parts of trajectory different weather prediction models could be used to provide this transformation based on actual weather data [27, 28].

At the second step of trajectory data process, the parameters of airplane trajectory could be estimated by the set of synchronized position coordinates [29]. This task usually is solved in some local cartesian reference frame like North-East-Up (NEU) coordinate system. A reference point of NEU is located at the body of reference ellipsoid (WGS84 model is commonly used). Z axis is tangential to the ellipsoidal surface and X is directed to the North. In the second step, parameters of trajectory are estimated: heading angles, ground and vertical velocities, and accelerations. In the case of real-time data processing, these parameters are used in the airplane model to predict the time of arrival at each upcoming waypoint of a flight plan.

Spline functions could be used for “filling the gaps” in the ADS-B datasets. A regression model with spline functions could be used for data interpolation in post-processing mode [30, 31]:

$$F=BC, \tag{1}$$

$$B = \begin{bmatrix} B_{1,m}(t_1) & \cdots & B_{k,m}(t_1) \\ \vdots & & \vdots \\ B_{1,m}(t_n) & \cdots & B_{k,m}(t_n) \end{bmatrix}, C = \begin{bmatrix} p_1 & t_1 \\ p_2 & t_2 \\ \cdots & \cdots \\ p_k & t_k \end{bmatrix},$$

where p is a parameter that is interpolated; C is the control points matrix; B is basic functions.

We propose to use the following form of basic function:

$$B_{j,1}(t) = \begin{cases} 1 \; if \; \tau_j \le t \le \tau_{j+1} \\ 0 \; if \; \tau_j > t > \tau_{j+1} \end{cases}$$

$$B_{j,m}(t) = \frac{t-\tau_j}{\tau_{j+m-1}-\tau_j} B_{j,m}(t) + \frac{\tau_{j+m}-t}{\tau_{j+m}-\tau_{j+1}} B_{j+1,m-1}(t), \tag{2}$$

where τ is a spline function node; *m, j* are order and basis function numbers.

Matrix of spline control points C is calculated with the help of available dataset X:

$$C = (B^T B)^{-1} B^T X. \tag{3}$$

$$X = \begin{bmatrix} x_1 & t_1 \\ x_2 & t_2 \\ \cdots & \cdots \\ x_k & t_k \end{bmatrix}.$$

Interpolated by (1) the full sequence of data points could be obtained for any frequency, as an example 1 HZ or higher. Matrix X could include four columns of latitude, longitude, altitude, and time. It makes possible to obtain results of trajectory data interpolation for each second in matrix F simultaneously.

III. Orientation Angles Estimation

Airplane orientation angles could be measured on-board only [32] or estimated based on known coordinates of each point of considered trajectory and weather data along airplane trajectory for the Date Time of airplane flight.

Orientation angles include angles of Roll, Pitch, and Yaw. These angles identify object (airplane) orientation at a particular point of airspace. Pitch angle orients airplane in a vertical plane based on a reference point in center mass. Yaw orients in a horizontal plane. In case, if Yaw angle is counted from the North direction clockwise it will coincide with the Heading angle [33]. However, yaw is a momentum angle from the previous value. Roll angle is a result of ailerons action and rotates along the longitudinal axis (front to back) of airplane body. Angles of pitch and heading (yaw) could be easily recovered from ADS-B data based on assumption of direct airplane movement on one side along the airplane trajectory.

Airplane trajectory in latitude (ϕ), longitude (λ), and altitude (h) should be transformed to ECEF:

$$P_{ECEF} = \begin{pmatrix} x_{ECEF} \\ y_{ECEF} \\ z_{ECEF} \end{pmatrix} = \begin{pmatrix} (N+h)\cos\phi\cos\lambda \\ (N+h)\cos\phi\sin\lambda \\ [N(1+e^2)+h]\sin\phi \end{pmatrix}, \tag{4}$$

$$N = a\left(1-e^2 sin^2(\phi)\right)^{-0.5},$$

where a is the equatorial radius of reference ellipsoid; e is eccentricity.

Transformation to local NED/NEU reference frame could be performed as follows:

$$P_{NED} = R\left(P_{ECEF} - P_{ECEF,ref}\right), \tag{5}$$

$$R = \begin{bmatrix} -\sin\phi_{ref}\cos\lambda_{ref} & -\sin\phi_{ref}\sin\lambda_{ref} & \cos\phi_{ref} \\ -\sin\lambda_{ref} & \cos\lambda_{ref} & 0 \\ -\cos\phi_{ref}\cos\lambda_{ref} & -\cos\phi_{ref}\sin\lambda_{ref} & \sin\phi_{ref} \end{bmatrix},$$

where $P_{ECEF,ref}$ is a reference point of NED system in ECEF; R is a transformation matrix from ECEF to NED; ϕ_{ref} and λ_{ref} are latitude and longitude of a reference point of NED system.

Pitch (θ) and heading (α) angles could be calculated based on geometrical relation presented in Fig. 2.

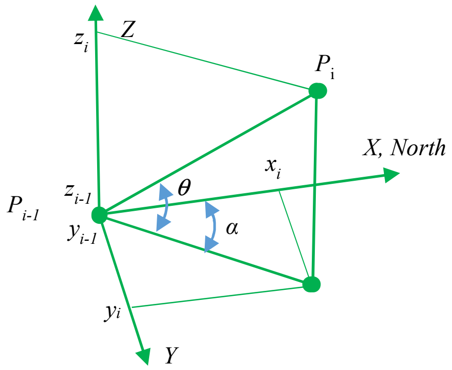


Fig. 2. Geometrical relation in trajectory data.

An algorithm for heading angle calculation is the following:

if $x_i - x_{i-1} > 0;\ y_i - y_{i-1} \ge 0$

$$\alpha = arctg\left(\frac{y_i - y_{i-1}}{x_i - x_{i-1}}\right);$$

if $x_i - x_{i-1} < 0; y_i - y_{i-1} \ge 0$

$$\alpha = \pi + arctg\left(\frac{y_i - y_{i-1}}{x_i - x_{i-1}}\right);$$

if $x_i - x_{i-1} < 0; y_i - y_{i-1} < 0$

$$\alpha = \pi + arctg\left(\frac{y_i - y_{i-1}}{x_i - x_{i-1}}\right);$$

if $x_i - x_{i-1} > 0; y_i - y_{i-1} < 0$

$$\alpha = 2\pi + arctg\left(\frac{y_i - y_{i-1}}{x_i - x_{i-1}}\right);$$

if $x_i - x_{i-1} = 0; y_i - y_{i-1} > 0$

$$\alpha = \frac{\pi}{2};$$

if $x_i - x_{i-1} = 0; y_i - y_{i-1} < 0$

$$\alpha = \frac{3\pi}{2}.$$

Pitch angle could be calculated as follows:

$$\theta = arctg\left(\frac{z_i - z_{i-1}}{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}}\right). \quad (6)$$

If data are interpolated for 1Hz (for each second $t_i - t_{i-1} = 1s$), then vertical velocity could be estimated as follows:

$$V_h = \frac{z_i - z_{i-1}}{t_i - t_{i-1}}. \quad (7)$$

Ground speed could be estimated as follows:

$$V_G = \frac{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}}{t_i - t_{i-1}}. \quad (8)$$

Total speed could be estimated as follows:

$$V_T = \frac{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2}}{t_i - t_{i-1}}. \quad (8)$$

In order to simplify calculation we move reference point of NEU coordinate system for each $(i - 1)$ point of trajectory data set during parameters calculation. It makes $x_{i-1} = 0$; $y_{i-1} = 0$; $z_{i-1} = 0$.

IV. Numerical Demonstration

As an example let's use ADS-B trajectory data set for UAE83 flight operated by May 15, 2023 between Dubai International (OMDB) and Geneva Cointrin International (LSGG) airports. The main parameters of trajectory data are presented in Table I. Trajectory dataset includes only 611 points. Interpolation by (1) gives 24720 data points (for each second of flight).

Results of heading and pitch angles calculation by (6) are represented in Fig. 3 and Fig. 4 correspondently at Eastern European Time scale. Results of velocity estimation by (7)-(9) are presented in Fig. 5 and Fig. 6.

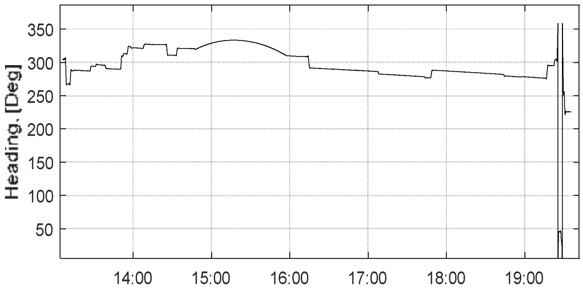


Fig. 3. Results of heading angle calculation.

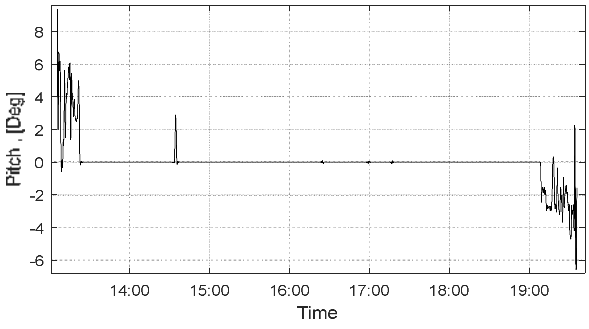


Fig. 4. Results of pitch angle calculation.

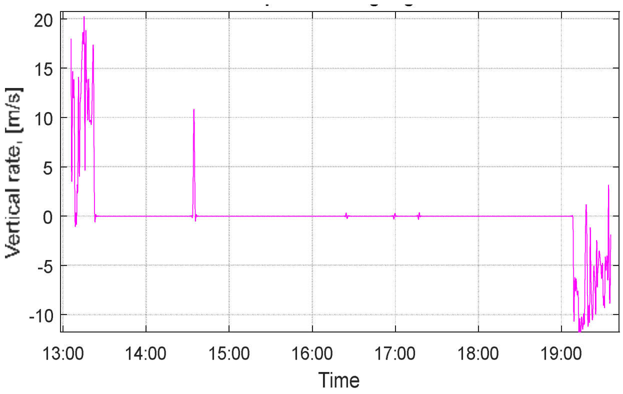


Fig. 5. Calctulated vertical rate.

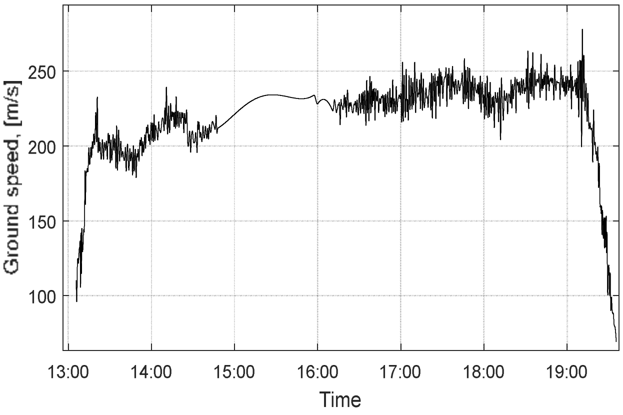


Fig. 6. Results of velocity calculation.

The whole flight mostly uses FL 320 as a cruise altitude. Thus a pitch angle is positive in the climbing process at the beginning and negative during descending at the destination airport vicinity. Also, pitch angle is strongly correlated with vertical velocity.

Conclusion

In the common case, the whole airplane trajectory obtained by ADS-B includes multiple gaps. Mostly these gaps are result of the poor coverage area of ground network of SDR. Data interpolation by linear regression model with spline functions gives a good result with simple formulas usage which requires low computation time. Proposed model (2) of B-splines basis function of second order guarantees continuity and smoothness of interpolated data that the best-fit input trajectory dataset. Also, usage of a spline model (3) gives the full set of interpolated parameters (latitude, longitude, and altitude) in one iteration step that helps to save computation power. Recovered the whole data set of a particular airplane trajectory for each second can be useful in surveillance data processing and analysis for ensuring the required level of flight safety.

Angles of pith and heading can be easily recovered from trajectory data based on proposed in paper formulas. Also, it should be noted that airplane orientation angles are tense to external action of weather. Thus for accurate orientation angles calculation, a wind direction and wind speed are required along the whole airplane trajectory. Unfortunately, the roll angle could not be estimated correctly based on trajectory data only.

Vertical, ground and total airplane speeds could be estimated based on trajectory data. Calculated by ADS-B data orientation angles and airplane velocity are tense to errors of on-board positioning system and pressure altimeter. In further studies, we will study error models of calculated orientation angles and velocities.

REFERENCES

[1] European ATM Master Plan. Digitalising Europe’s Aviation Infrastructure. SESAR. 2020.

[2] S. Zhyla, V. Volosyuk, V. Pavlikov, N. Ruzhentsev, E. Tserne, and A. Popov, “Statistical synthesis of aerospace radars structure with optimal spatio-temporal signal processing, extended observation area and high spatial resolution,” in Radioelectronic and computer systems, vol. 101, issue 1, 2022, pp. 178–194.

[3] N. Ruzhentsev, S. Zhyla, V. Pavlikov, V. Volosyuk, E. Tserne, A. Popov, “Radio-Heat Contrasts of UAVs and Their Weather Variability at 12 GHz, 20 GHz, 34 GHz, and 94 GHz Frequencies,” in ECTI Transactions on Electrical Engineering, Electronics, and Communications, vol 20, issue 2, 2022, pp. 163–173.

[4] I.V. Ostroumov and N.S. Kuzmenko, “Statistical Analysis and Flight Route Extraction from Automatic Dependent Surveillance-Broadcast Data,” 2022 Integrated Communications Navigation and Surveillance Conference (ICNS), April 2022, pp. 1–9.

[5] M. Shravan, R. Rakshit, P. Sanjana, B. K. Priya, and N. Kumar, “RTL SDR ADS-B Data Analysis for Predicting Airports and ATS Routes,” in International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1–7.

[6] F. Eichstaedt and J. Budroweit, “Evaluation of a GNURadio-based multi-channel ADS-B receiver implemented on a highly integrated SDR platform for space application,” in IEEE Space Hardware and Radio Conference (SHaRC), Las Vegas, NV, USA, 2022, pp. 11–14.

[7] O. Solomentsev, M. Zaliskyi, and O. Zuiev, “Radioelectronic Equipment Availability Factor Models,” Signal Processing Symposium 2013 (SPS 2013), Serock, Poland, 2013, pp. 1–4.

[8] O. Solomentsev, M. Zaliskyi, Y. Averyanova, N. Kuzmenko, B. Kuznetsov, and T. Nikitina, “Method of Optimal Threshold Calculation in Case of Radio Equipment Maintenance,” in Data Science and Security. Lecture Notes in Networks and Systems, vol. 462, 2022, Springer, Singapore, pp. 69–79.

[9] O.C. Okoro, M. Zaliskyi, S. Dmytriiev, O. Solomentsev, and O. Sribna, “Optimization of Maintenance Task Interval of Aircraft Systems,” International Journal of Computer Network and Information Security (IJCNIS), 2022, vol.14, No. 2, pp. 77–89.

[10] O. Solomentsev, M. Zaliskyi, T. Herasymenko, O. Kozhokhina and Y. Petrova, “Data Processing in Case of Radio Equipment Reliability Parameters Monitoring,” in Advances in Wireless and Optical Communications (RTUWO), 2018, pp. 219–222.

[11] N. Pearce, K. J. Duncan, and B. Jonas, “Signal Discrimination and Exploitation of ADS-B Transmission,” in SoutheastCon 2021, Atlanta, GA, USA, 2021, pp. 1–4.

[12] O. Sushchenko, Y. Averyanova, M. Zaliskyi, O. Solomentsev, B. Kuznetsov, and T. Nikitina, “Algorithms for Design of Robust Stabilization Systems,” in Computational Science and Its Applications – ICCSA 2022. Lecture Notes in Computer Science, vol.13375, 2022, Springer, Cham, pp. 198–213.

[13] O. Sushchenko, Y. Bezkorovainyi, V. Golitsyn, and N. Kuzmenko, “Integration of MEMS Inertial and Magnetic Field Sensors for Tracking Power Lines,” In XVIII International Conference on the Perspective Technologies and Methods in MEMS Design (MEMSTECH), 2022, pp. 33–36.

[14] I. Ostroumov, N. Kuzmenko, and Y. Bezkorovainyi, “Relative navigation for vehicle formation movement,” In 3rd KhPI Week on Advanced Technology, Kharkiv, Ukraine, Oct. 03, 2022, pp. 10–13.

[15] I.V. Ostroumov and N.S. Kuzmenko, “An area navigation RNAV system performance monitoring and alerting,” in 1st International Conference System Analysis & Intelligent Computing: SAIC 2018, IEEE, October 2018, pp. 211–214.

[16] T. M. Adami and J. J. Zhu, “6DOF flight control of fixed-wing aircraft by Trajectory Linearization,” in American Control Conference, San Francisco, CA, USA, 2011, pp. 1610–1617.

[17] N. Kuzmenko, I. Ostroumov, Y. Bezkorovainyi, Y. Averyanova, and V. Larin, “Airplane Flight Phase Identification Using Maximum Posterior Probability Method,” 3rd International Conference on System Analysis & Intelligent Computing (SAIC),Kyiv, Ukraine, October 4-7, 2022, pp. 1-5,

[18] D. Coiro, A. De Marco, and F. Nicolosi, “A 6DOF flight simulation environment for general aviation aircraft with control loading reproduction,” in AIAA Modeling and Simulation Technologies Conference and Exhibit, 2007, p. 6364.

[19] Z. Yongguo, H. Xiang, F. Wei, and L. Shuanggao, “Trajectory planning algorithm based on quaternion for 6-DOF aircraft wing automatic position and pose adjustment method,” Chinese Journal of Aeronautics, vol. 23, issue 6, 2010, pp. 707–714.

[20] C.S. Jamadagni, C.U. Chethan, Y. Jeppu, S.B. Kamble, and V.H. Desai, “System simulation approach for helicopter autopilot,” in International Conference on Contemporary Computing and Informatics (IC3I), Mysore, India, 2014, pp. 404–408.

[21] X. Zhu, X. Li, and X. Gong, “Performance Analysis of ADS-B Receiving System Based on LEO Satellite Constellation,” in 21st International Symposium on Communications and Information Technologies (ISCIT), 2022, pp. 39–42.

[22] J. Sun, The 1090 megahertz riddle: a guide to decoding mode S and ADS-B signals. TU Delft OPEN Publishing, 2021.

[23] Technical Provisions for Mode S Services and Extended Squitter, 2008, Doc 9871, First Edition, Monreal, Canada, ICAO.

[24] M. Zaliskyi, O. Solomentsev, V. Larin, and Y. Averyanova, “Model Building for Diagnostic Variables during Aviation Equipment Maintenance,” In 17th International Conference on Computer Sciences and Information Technologies (CSIT), 2022, pp. 160–164.

[25] O. Sushchenko, Y. Bezkorovainyi, and Y. Averyanova, “Data Processing through the Lifecycle of Aviation Radio Equipment,” In 17th International Conference on Computer Sciences and Information Technologies (CSIT), IEEE, 2022, pp. 146–151.

[26] V. Larin, O. Solomentsev, M. Zaliskyi, and A. Shcherban, “Prediction of the final discharge of the UAV battery based on fuzzy logic estimation of information and influencing parameters,” 3rd KhPI Week on Advanced Technology (KhPI Week), Kharkiv, Ukraine, October 03-07, 2022, pp.44–49.

[27] Y. Averyanova, V. Larin, N. Kuzmenko, and O. Solomentsev, “Turbulence Detection and Classification Algorithm Using Data from AWR,” in 2nd Ukrainian Microwave Week (UkrMW), IEEE, Ukraine, 2022, pp. 518–522.

[28] A. Popov, E. Tserne, S. Zhyla, V. Volosyuk, V. Pavlikov, and N. Ruzhentsev, “Invariant polarization signatures for recognition of hydrometeors by airborne weather radars,” in. Computational Science and Its Applications – ICCSA 2023. Lecture Notes in Computer Science, vol.13956, 2023, Springer, Cham, pp. 1–14.

[29] S. Zhyla, V. Volosyuk, V. Pavlikov, N. Ruzhentsev, E. Tserne, and A. Popov, “Practical imaging algorithms in ultra-wideband radar systems using active aperture synthesis and stochastic probing signals,” in Radioelectronic and computer systems, vol.105, issue 1, 2023, pp. 55-73.

[30] I.V. Ostroumov, K. Marais, and N.S. Kuzmenko, “Aircraft positioning using multiple distance measurements and spline prediction,” in Aviation, vol. 26, issue 1, 2022, pp. 1–10.

[31] I.V. Ostroumov and N.S. Kuzmenko, “Accuracy improvement of VOR/VOR navigation with angle extrapolation by linear regression,” in Telecommunications and Radio Engineering, vol. 78, issue 15, 2019, pp. 1399–1412.

[32] K. Dergachov, O. Havrylenko, V. Pavlikov, A. Popov, and S. Zhyla, “GPS Usage Analysis for Angular Orientation Practical Tasks Solving,” in IEEE International Conference on Problems of Infocommunications. Science and Technology, Kyiv, 2022, pp. 1–6.

[33] O. Sushchenko, Y. Bezkorovainyi, O. Solomentsev, and N. Kuzmenko “Airborne Sensor for Measuring Components of Terrestrial Magnetic Field,” in 41st International Conference on Electronics and Nanotechnology (ELNANO), IEEE, Kyiv, Ukraine, October 10-14, 2022, pp. 687–691.

Performance Analysis of Compact Position Report for Geodata Storing and Transfering

Ivan Ostroumov
Air Navigation Systems Department
National Aviation University
Kyiv, Ukraine
ostroumovv@ukr.net

Olha Sushchenko
Aerospace Control Systems Department
National Aviation University
Kyiv, Ukraine
sushoa@ukr.net

Yuliya Averyanova
Air Navigation Systems Department
National Aviation University
Kyiv, Ukraine
ayua@nau.edu.ua

Maksym Zaliskyi
Telecommunication and
Radioelectronic Systems Department
National Aviation University
Kyiv, Ukraine
maximus2812@ukr.net

Oleksandr Solomentsev
Telecommunication and
Radioelectronic Systems Department
National Aviation University
Kyiv, Ukraine
avsolomentsev@ukr.net

Oleksii Holubnychyi
Telecommunication and
Radioelectronic Systems Department
National Aviation University
Kyiv, Ukraine
oleksii.holubnychyi@npp.nau.edu.ua

Abstract — **Nowadays geographic data (geodata) are used in a variety of applications. A compact position report (CPR) is one of the commonly used algorithms for reducing the size of geographic coordinates, which are represented in form of latitude and longitude. CPR grounds on automatic scaling technology which adobes the size of input data to the number of available bits for data storage. In the paper, we study adaptive bit selection in CPR based on precision of input data. The area of uncertainty rectangle and standard deviation error of positioning in horizontal plane are used to analyze the optimal number of bits. A confidence band of 99.7% is used to estimate error ellipsoid for geographic position holding. Also, performance of CPR algorithm is studied for different levels of input data precision.**

***Keywords* — geodata, compact position report, position, latitude, longitude, performance, coding, tracking, precision**

I. Introduction

Geographic data are widely used today in different applications [1, 2]. Geographic data (Geodata) is associated with a particular point of location or affected area in space. Geodata is used in geodesy for precise location measurements of a group of objects, as well as different parameter measurements by vary of mobile sensors [3, 4]. Common cameras use geodata to put in output file descriptions of videos or pictures taken. Geodata helps to know exactly location of services and goods [5]. Most search engines use geodata to provide geo-oriented search of requited data, which gives more weight to services placed close to user location.

Geodata plays a special place in the tracking of vehicles and moving objects in space [6]. It connects with continuously changed coordinates of object location in space. Passenger and cargo transportation use multiple geodata at different levels of vehicle control and monitoring [7, 8]. In air transportation, geodata is used for air traffic control, navigation, and surveillance to ensure the required level of flight safety [9]. Surveillance is supported with a different type of radars [10, 11], which measure coordinates of each airspace user together with weather data [12]. Modern concept of Air Traffic Management requires each airspace user to be equipped with a specific transponder to share its position with all other users in open data format [13, 14]. On-board vehicle transponder of Automatic Dependent Surveillance-Broadcast (ADS-B) transmits vehicle identification and position reports with geodata [15, 16].

Surveillance according to ADS-B concept considers measuring vehicle position by on-board sensors and sharing coordinates with a help of digital data link for any other airspace user omnidirectionally [17, 18]. ADS-B transponder is required for each airspace user and ground vehicle operated on the runway or taxiways of airport facilities. Ground implementation of ADS-B works the same as on-board including maintaining and operational data processing [19], [20]. Transponder of Mode 1090ES is one of the most useful ADS-B equipment worldwide.

The most useful coordinate system for global geodata representation is Latitude, Longitude, and Altitude (LLA). LLA perfectly works for object localization near the surface of the Earth's ellipsoidal model. In most cases, angles of latitude and longitude are used for object localization in the surface of elipsoidal model (or projections on a horizontal plane) and altitude for Three Dimensional space. The latitude range of 180° and longitude of 360° are used to point objects on the global ellipsoidal surface. In common, accurate positioning required to use of 5 digits after a comma to guarantee position precision in a few meters. Most vehicles used today are moved in the local area only. Automotive, railway, maritime, and air transportation vehicles most time operate in the local area with low speed (in comparison to changing the whole range of coordinates) [21, 22]. Also, one-trip tracking does not require to use the whole range of latitude and longitude. In this case, it does not make any sense to use the whole range of latitude and longitude to describe vehicle movement in order to save and archive trajectory data. Amount of space on server equipment for data storage directly depends on the number of bits required to hold digits of latitude and longitude coordinates.

A compact position report (CPR) is used to reduce the size of space for coordinates data transferring and storage. There are different types of CPR algorithms [23, 24]. Common CPR grounds on division of latitude and longitude whole range into specific zones and applying coding for angles within particular zone by numbers with guided scale [25]. As an example, data transferred in ADS-B are coded by CPR which helps to reduce number of bits from required 45 (for the case of transmission data without coding) to 34 bits. In this case, CPR saves 11 bits in each digital message with no one negative impact on the precision of transmitted data.

In the paper, we study performance of CPR algorithm and analyze accuracy reduction based on performance of input

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

data and the size of bits available for data storage. For the last decades, performance of positioning sensors is improved significantly. Therefore, the study of CPR algorithm operation with different precision level data is an important step to improve performance of tracking and geodata storage systems.

II. Compact Position Reporting Algorithm

CPR is a multi-parameter coding and can be used for any sort of data. However, CPR is commonly used for operations with geodata in different tracking system designs. CPR divides the whole range for latitude and longitude into multiple small zones, and then codes data by coordinates in the local reference frame by a particular scale within a selected zone (Fig. 1). Note, that the local reference frame is not cartesian due to axes direction parallel to meridians and parallels. CPR does not use any information about zone. Zone number is identified automatically during decoding procedure. There are two basic types of zone identification: usage of a different number of zones and sequential zone number tracking. The first one uses two different amount of zones (usually called odd and even). Odd zone number using one zone less from even coding. As an example, CPR in ADS-B uses 60 zones for even and 59 for odd [26].

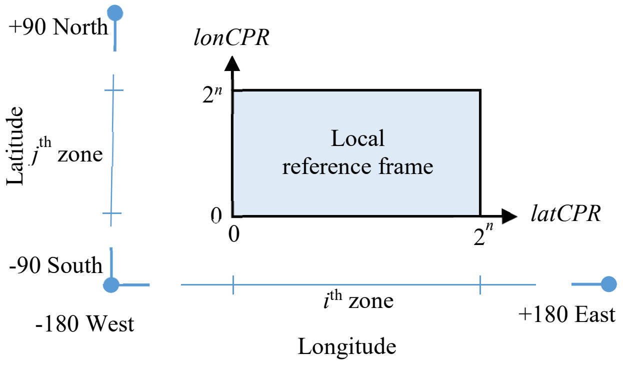


Fig. 1. Coding data by CPR.

During the decoding results of odd and even data are compared for each global rectangle. An actual zone is identified by minimum distance between even and odd decoded data (Fig. 2). Another approach is grounded on vehicle trajectory tracking and in case of vehicle operation near the perils of a rectangle such algorithms consider switching to neighbor zone by the closed distance between a precious position inside of known zone and current position with neighbor zones [27]. Thus, saving space in CPR is a result of using local reference frame which is associated with a particular rectangle.

Geodata uses coordinates in LLA. Latitude is an angle between a line connected with user location on the ellipsoidal surface and equatorial plane. Latitude is in ranges from –90° at the South pole to 90° at the North pole, with 0° at the Equator. Zones of constant latitude create parallels, which are circles located parallel to the equatorial plane.

Longitude is an angle between a line connected with user location and the prime meridian plane. Longitude is in ranges from –180°W to 180°E, with 0° at the Prime meridian. The number of longitude zones is different for each latitude zones, which reaches its maximum on the equatorial and minimum on the poles (Fig. 3).

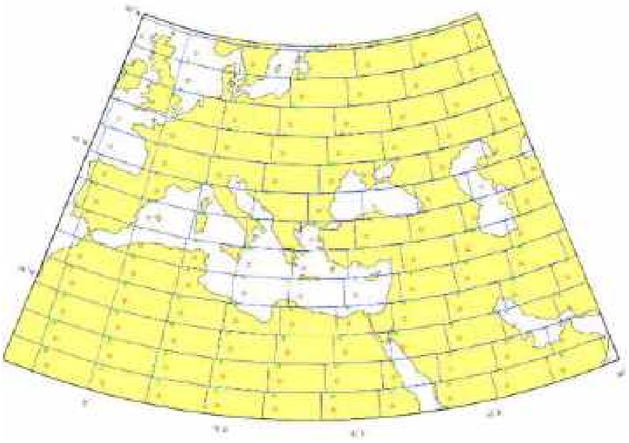


Fig. 2. Zone identification by distance between even and odd coded data.

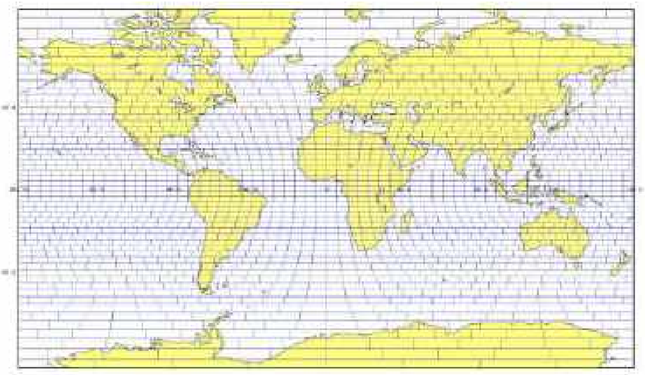


Fig. 3. Zones of longitude and latitude in CPR.

Rectangle specified by a particular zone of latitude and longitude is used as a local reference frame with a particular scale for CPR coding. It uses a scale with 0 minimal limit and 2n is an upper peril, where *n* is the number of bits available for coded value storage.

Geodata in CPR is coded by fooling equation [28]:

$$latCPR = mod\left(floor\left(2^n \frac{mod(lat, \Delta lat)}{\Delta lat} + 0{,}5\right), 2^n\right), \quad (1)$$

$$lonCPR = mod\left(floor\left(2^n \frac{mod(lon, \Delta lon)}{\Delta lon} + 0{,}5\right), 2^n\right), \quad (2)$$

where *lat* and *lon* are input latitude and longitude of user position in degrees; *Δlat* and *Δlon* are widths of latitude and longitudinal zones used; *n* is a number of bits used for data coding; *mod()* is function that returns the remainder after division of first value into the second one; *floor()* is a function which rounds to the nearest integer less than or equal to input value.

The width of each zone is calculated by division of whole range of latitude and longitude into the number of zones used:

$$\Delta lat = \frac{180}{N}, \quad (3)$$

$$\Delta lon = \frac{360}{max(1, NL(lat))}, \quad (4)$$

where *N* is the number of zones on latitude side; *NL()* is the number of longitudinal zones for a particular zone of latitude.

The number of zones in a longitudinal direction is calculated for each latitude zone separately as follows [29]:

$$NL(lat) = floor\left(\frac{2\pi}{acos(A)}\right). \quad (5)$$

$$A = 1 - \frac{1-cos\left(\frac{\pi}{N}\right)}{cos^2\left(\frac{\pi\, lat}{180}\right)}. \quad (6)$$

The maximum number of zones in longitudinal direction makes sense to use in double times bigger than in latitude one, due to the scale sizes proportion, latitude uses 180° and longitude uses 360°.

Geodata coded by (1) and (2) is automatically zoomed to available amount of bits *n*. However, in the case, when number of bits is much lower than required, the precision of data transferring is reduced.

III. Performance of CPR

Performance of coded by CPR geodata depends on precision of input latitude and longitude measurements [30, 31] and accuracy degradation caused by not correct number of bits used. A scale of local reference frame specifies resolution of CPR. Resolution of CPR can be represented as an area or elementary rectangle or by axes of ellipse with a rectangle inside (Fig. 4).

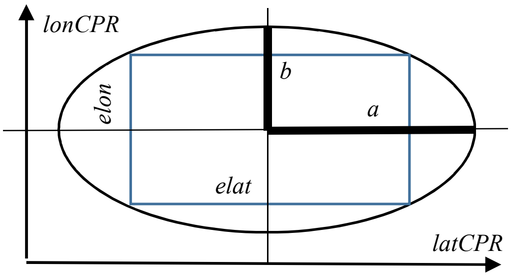


Fig. 4. Uncertainty area of CPR.

The sides of an elementary rectangle can be estimated based on arch length as follows:

$$elon = R\frac{\Delta lon}{2^n}, \; elat = R\frac{\Delta lat}{2^n}, \quad (7)$$

where *elon* is the length of elementary side in the North-South direction; *elat* is the length of elementary side in the East-West direction; *Δlon* and *Δlat* are widths of longitude and latitude zones in radians; *R* is Earth radius (global average value is 6371 km).

Uncertainty area can be calculated as follows:

$$Su = elon\; elat = R^2\frac{\Delta lon \Delta lat}{2^n}. \quad (8)$$

Precision of CPR also can be analyzed with the help standard deviation of positioning. Uncertainty rectangle can be represented inside of uncertainty ellipse (Fig. 4). In this case, the smallest ellipse which includes an uncertainty rectangle is used. Semi-axes of uncertainty ellipse can be calculated from the known height and length of rectangle, as follows:

$$b = \frac{elon\sqrt{2}}{2}, \; a = \frac{elat\sqrt{2}}{2}, \quad (9)$$

where *b* is the semi-minor axis; *a* is the semi-major axis.

Due to exactly known user location in the uncertainty rectangle, it makes sense to assume that most data are inside of uncertainty ellipse with a 99.7% length of confidence band. Based on the “3 Sigma” rule, standard deviations can be estimated as follows:

$$\sigma_x = \frac{a}{3} = \frac{elat\sqrt{2}}{6}, \; \sigma_y = \frac{b}{3} = \frac{elon\sqrt{2}}{6}, \quad (10)$$

$$\sigma_p = \sqrt{\sigma_x^2 + \sigma_y^2} = \sqrt{\frac{elat^2 + elon^2}{18}}, \quad (11)$$

where σ_x is the standard error deviation in the North-East direction; σ_y is the standard error deviation in the West-East direction.

A simple analysis of equation (11) gives that size of elementary rectangle has a direct influence on positioning performance. Therefore, in order to improve CPR performance, it is important to reduce values *elat* and *elon*.

Another important component is the precision of input data, which means a resolution of measured data and accuracy of positioning sensor. Due to measurements in latitude and longitude being fixed on a cyclic scale, let’s analyze the uncertainty level caused by each precision digit. We use precision of latitude and longitude in digits of degrees after point to calculate the area of uncertainty rectangle by (8) and standard deviation of positioning error by (11). Result of these calculations is represented in Table I.

TABLE I. Analysis of Input Data Precision

No	Precision level, [deg]	Area of uncertainty rectangle, [km^2]	$\sigma_{p,}$ [m]	Approximate level of data precision
1	10^{-1}	123.6	1235.5	1 km
2	10^{-2}	1.23	123.5	100 m
3	10^{-3}	0.012	12.35	10 m
2	10^{-4}	1.2×10^{-4}	1.235	1 m
5	10^{-5}	1.2×10^{-6}	0.135	1 dm
6	10^{-6}	1.2×10^{-8}	0.0124	1 sm
7	10^{-7}	1.2×10^{-10}	0.0012	1 mm

Obtained result indicate that coordinate data in format with four digits after point gives approximately 1 m precision, five digits – 1 dm, six digits – 1 sm, seven – 1 mm. Also, results of uncertainty area and standard deviation are presented in Fig. 5.

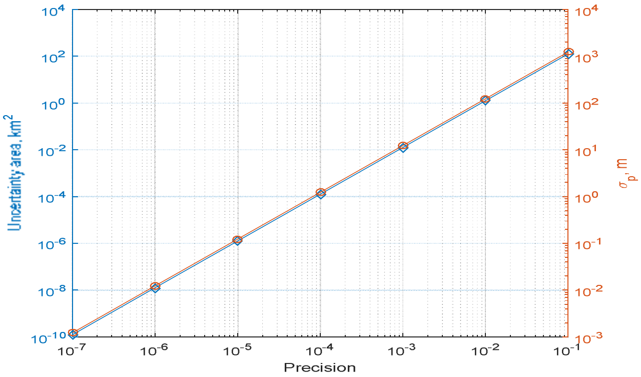


Fig. 5. Precision of input to CPR data.

During selection of CPR settings, it is important to take into account the uncertainty area and standard deviation of input data for effective space usage. Values of input precision level should coincide with a precision of CPR.

IV. Effective CPR Coding

Let's analyze effective CPR coding for vehicle position measured in coordinates in latitude and longitude with 1 dm precision (five digits after coma), which is a common performance of multi-constellation global navigation satellite system for stationary object positioning.

Results of the analysis uncertainty area for the range of total zones number from 30 to 100 and an available number of bits estimated by (8) are represented in Fig. 6.

Results of standard deviation error analysis by (11) are presented in Fig. 7. Values of σ_P for range from 0.1 to 0.9 are required to meet requirements of 1 dm precision based on Table I.

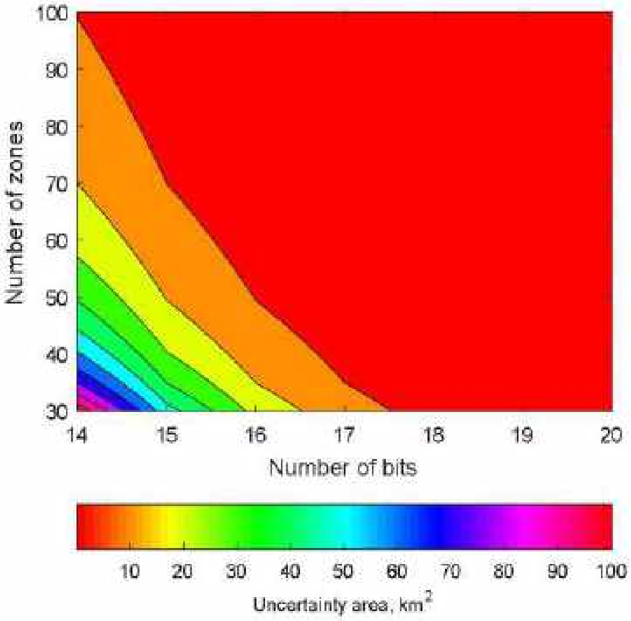


Fig. 6. Uncertainty area for CPR with different settings.

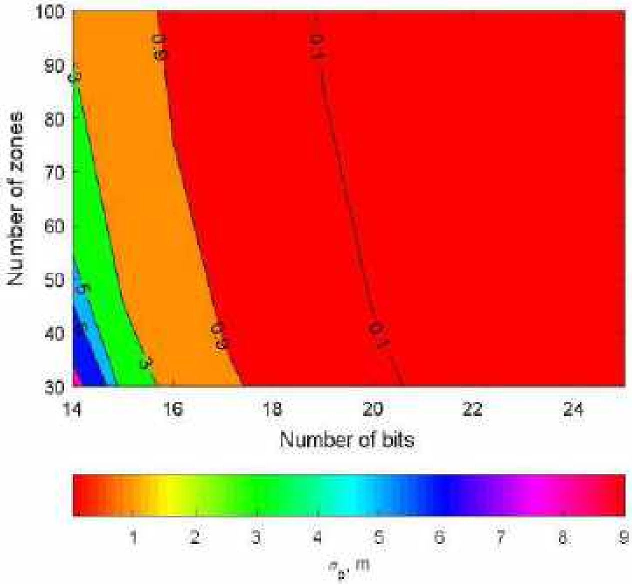


Fig. 7. Standard deviation.

Obtained results indicate that 18 bits of data are required for CPR coding of 1 dm precision of input coordinates without accuracy degradation. Results of CPR performance analysis for different precision levels, indicated in Table I, are represented in Fig. 8.

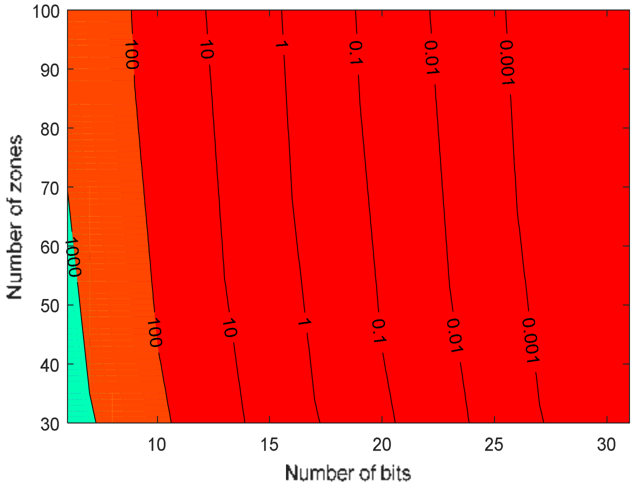


Fig. 8. Standard deviation error correspondence to the precision level.

Taking into account that the maximum data of longitude is 359º plus precision level, it is possible to analyze performance of CPR by the number of saved bits after applying codding. Results of common CPR coding performance is represented in Table II.

TABLE II. CPR Performance

No	Precision level, [deg]	Minimal number of bits		
		Raw data	*CPR*	*Saved number of bits*
1	10^{-1}	12	8	5
2	10^{-2}	16	11	5
3	10^{-3}	19	14	5
2	10^{-4}	22	17	5
5	10^{-5}	26	21	5
6	10^{-6}	29	24	5
7	10^{-7}	32	27	5

Raw data precision has been estimated as follows:

$$PL = round(log_2(M)), \qquad (12)$$

where M is a maximum number in degrees ignoring float point.

Obtained results indicate that CPR coding gives approximately the same space-saving capability for different precision levels. Five bits saving in each coordinate are double for the entire one-point position of geodata.

Conclusions

CPR is an important component of geodata representation. Modern algorithms of CPR coding are adaptive to space, which is available for geodata holding. However, effective CPR coding requires correspondence of CPR performance and accuracy of input coordinates.

Precision of input latitude and longitude data specify general CPR settings including number of bits required for data transition without accuracy degradation. Variety of CPR applications with different types of sensors used require segmentation correspondent size for CPR-coded data storage.

Obtained results indicate that the decimeter precision of geodata requires at least 21 bits of data for CPR coding. In comparison with using raw data transmission of one coordinate for serve 359 degrees with five numbers after point requires at least 26 bits. Thus, CPR saves 5 bits for each coordinate, which gives saving 10 bits for storage of one point of geodata.

References

[1] M. Helbich, C. Amelunxen, P. Neis, and A. Zipf, “Comparative spatial analysis of positional accuracy of OpenStreetMap and proprietary geodata,” Proceedings of GI_Forum, 4, 2012, pp. 24–33.

[2] G. Qiao, W. Wang, Z. Wu, and J. Zhang, “Web service research of urban geographical data based on xml,” In 6th International Conference on ITS Telecommunications IEER, 2006, pp. 214–217.

[3] T. Pei, C. Song, S. Guo, H. Shu, Y. Liu, Y. Du, and C. Zhou, “Big geodata mining: Objective, connotations and research issues,” Journal of Geographical Sciences, vol. 30(2), 2020, pp. 251–266.

[4] M. H. Tsou and B. P. Buttenfield, “A dynamic architecture for distributing geographic information services,” Transactions in GIS, vol. 6(4), 2002, pp. 355–381.

[5] E. Klien and M. Lutz, “The role of spatial relations in automating the semantic annotation of geodata,” In International Conference on Spatial Information Theory. Springer, Berlin, 2005, pp. 133–148.

[6] B. E. Mikkelsen, A. K. Lyseen, M. Dobroczynski, and H. S. Hansen, “Behavioural nutrition & big data: How geodata, register data & GPS, mobile positioning, Wi-Fi, Bluetooth & thermal cameras can contribute to the study of human food behaviour,” In Proceedings of the Measuring Behavior, 2014, pp. 285–299.

[7] I.V. Ostroumov and N.S. Kuzmenko, “An area navigation RNAV system performance monitoring and alerting,” in 1st International Conference of System Analysis & Intelligent Computing (SAIC), IEEE, 2018, pp. 211–214.

[8] M. Dzunda, P. Dzurovciv, I. Koblen, S. Szabo, E. Jenkova, P. Cekan, et. al., “Selected aspects of navigation system synthesis for increased flight safety, protection of human lives, and health,” in International Journal of Environmental Research and Public Health, vol. 17, issue 5, 2020, p. 1550. doi: 10.3390/ijerph17051550.

[9] M. Dzunda and P. Dzurovcin, “Influence of mutual position of communication network users on accuracy of positioning by telemetry method,” TransNav: International Journal on Marine Navigation and Safety of Sea Transportation, vol. 15, issue 2, 2021, pp. 299–306.

[10] S. Zhyla, V. Volosyuk, V. Pavlikov, N. Ruzhentsev, E. Tserne, and A. Popov, “Practical imaging algorithms in ultra-wideband radar systems using active aperture synthesis and stochastic probing signals,” in Radioelectronic and computer systems, vol.105, issue 1, 2023, pp. 55–73.

[11] V. Volosyuk, S. Zhyla, V. Pavlikov, N. Ruzhentsev, E. Tserne, and A. Popov, “Optimal Method for Polarization Selection of Stationary Objects Against the Background of the Earth’s Surface,” in International Journal of Electronics and Telecommunications, vol 68, issue. 1, 2022, pp. 83–89.

[12] S. Zhyla, V. Volosyuk, V. Pavlikov, N. Ruzhentsev, E. Tserne, and A. Popov, “Statistical synthesis of aerospace radars structure with optimal spatio-temporal signal processing, extended observation area and high spatial resolution,” in Radioelectronic and computer systems, vol.101, issue 1, 2022, pp. 178–194.

[13] A. Popov, E. Tserne, S. Zhyla, V. Volosyuk, V. Pavlikov, and N. Ruzhentsev, “Invariant polarization signatures for recognition of hydrometeors by airborne weather radars,” in. Computational Science and Its Applications – ICCSA 2023. Lecture Notes in Computer Science, vol.13956, 2023, Springer, Cham, pp. 1–14.

[14] Technical Provisions for Mode S Services and Extended Squitter, Doc. 9871, First Edition, Monreal, Canada, ICAO, 2008.

[15] European ATM Master Plan. Digitalising Europe’s Aviation Infrastructure, SESAR Joint Undertaking, Eurocontrol, 2020.

[16] K. Dergachov, O. Havrylenko, V. Pavlikov, A. Popov, and S. Zhyla, “GPS Usage Analysis for Angular Orientation Practical Tasks Solving,” 2022 IEEE International Conference on Problems of Infocommunications. Science and Technology, Kyiv, Ukraine, 2022, pp. 1–6.

[17] I.V. Ostroumov and N.S. Kuzmenko, “Statistical Analysis and Flight Route Extraction from Automatic Dependent Surveillance-Broadcast Data,” In Integrated Communications Navigation and Surveillance Conference (ICNS), April 2022, pp. 1–9.

[18] X. Zhu, X. Li, and X. Gong, “Performance Analysis of ADS-B Receiving System Based on LEO Satellite Constellation,” In 21st International Symposium on Communications and Information Technologies (ISCIT), 2022, pp. 39–42.

[19] X. Su, J. Li, J. Zeng, and S. Hu, “SDR Reception and Analysis of Civil Aviation ADS-B Signals,” In International Conference of Safety Produce Informatization (IICSPI), IEEE, 2018, pp. 893–896, doi: 10.1109/IICSPI.2018.8690376.

[20] M. Zaliskyi, O. Solomentsev, V. Larin, and Y. Averyanova, “Model Building for Diagnostic Variables during Aviation Equipment Maintenance,” In 17th International Conference on Computer Sciences and Information Technologies (CSIT), 2022, pp. 160–164.

[21] A. Syd, W. Schuster, W. Ochieng, A. Majumdar, “Analysis of anomalies in ADS-B and its GPS data. GPS solutions, vol. 20, 2016, pp. 429–438.

[22] I. Ostroumov, N. Kuzmenko, and Y. Bezkorovainyi, “Relative navigation for vehicle formation movement,” In 3rd KhPI Week on Advanced Technology (KhPI Week), Kharkiv, Ukraine, October 03-07, 2022, pp. 10–13.

[23] V. Larin, O. Solomentsev, A. Shcherban, and Y. Averyanova, “Prediction of the final discharge of the UAV battery based on fuzzy logic estimation of information and influencing parameters,” In 3rd KhPI Week on Advanced Technology (KhPI Week), Kharkiv, Ukraine, October 03-07, 2022, pp.44–49.

[24] L. Titolo, M. Moscato, C. A. Munoz, A. Dutle, and F. Bobot, “A formally verified floating-point implementation of the compact position reporting algorithm,” In Formal Methods: 22nd International Symposium, Held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 15-17, 2018, pp. 364–381.

[25] X. Haifei, W. Fei, Z. Xiaobo, Z. Haiqing and L. Yintian, “Dynamic Capture Position and Compression Algorithm in Logistics and Transportation Area,” In Third International Conference on Intelligent System Design and Engineering Applications, Hong Kong, China, 2013, pp. 622–625, doi: 10.1109/ISDEA.2012.150.

[26] A. Marshall, An expanded description of the CPR algorithm. Sensis Corporation, New York, 2009.

[27] O. Havrylenko K. Dergachov, V. Pavlikov, S. Zhyla, and O. Shmatko, “Decision Support System Based on the ELECTRE Method,” in Data Science and Security. Lecture Notes in Networks and Systems, vol. 462, 2022, Springer, Singapore, pp. 295–304.

[28] F. Musmann, “ADS-B and Functions for Flight Inspection,” In Proceedings of the 2018 International Flight Inspection Symposium, Monterey, California, April 16-20, 2018, pp. 111–123.

[29] A. Dutle, M. Moscato, C. Munoz, G. Anderson, and F. Bobot, “Formal analysis of the compact position reporting algorithm,” in Formal Aspects of Computing, vol. 33 issue 1, 2021, pp.65–86.

[30] J. Sun, “The 1090 megahertz riddle: a guide to decoding mode S and ADS-B signals,” TU Delft OPEN Publishing, 2021.

[31] I.V. Ostroumov, V.P. Kharchenko, and N.S. Kuzmenko, “An airspace analysis according to area navigation requirements,” in Aviation, vol. 23, issue 2, 2019, pp. 36-42.

Algorithm for Selecting a Surface Model for Remote Sensing of Earth's Surfaces

Kseniia Nezhalska
Aerospace Radioelectronic Systems Department
National Aerospace University "Kharkiv Aviation Institute"
Kharkiv, Ukraine
k.nezhalska@khai.edu

Konstantin Belousov
Spacecraft, Measuring Systems and Telecommunications Department
Yuzhnoye SDO
Dnipro, Ukraine
https://orcid.org/0000-0002-6436-3359

Valery Volosyuk
Aerospace Radioelectronic Systems Department
National Aerospace University "Kharkiv Aviation Institute"
Kharkiv, Ukraine
https://orcid.org/0000-0002-1442-6235

Semen Zhyla
Aerospace Radio-Electronic Systems Department
National Aerospace University “Kharkiv Aviation Institute”
Kharkiv, Ukraine
https://orcid.org/0000-0003-2989-8988

Oleksandr Mazurenko
Aerospace Radio-Electronic Systems Department
National Aerospace University “Kharkiv Aviation Institute”
Kharkiv, Ukraine
o.mazurenko@khai.edu

***Abstract* — In modern science and technology, remote studies and measurements are one of the main methods of analysis in many fields: natural sciences, national economy, climatology and others. At the same time, models of the surfaces are necessary tools that connect the parameters of this surfaces or objects with the electromagnetic fields that these surfaces emit or reflect. At present, in order to carry out practical measurements by remote sensing methods or simulation modeling of such experiments, first of all, it is necessary to choose a correct model of surface radiation or scattering. This stage of research is labor-consuming and time-consuming because of the necessity to go through a large number of model variants, to analyze the conditions of their application and relations between the parameters of measurement systems and characteristics of surfaces. This paper proposes a classification of surface models, a model selection algorithm, and a prototype of a software product for automating the process of selecting the relationship between the parameters of the investigated surface and the signals that are registered.**

Keywords — remote sensing, electrodynamic surface models, empirical surface models, radar measurements, radiometers

I. Introduction

Active [1] and passive [2] radioengineering remote sensing systems are currently used to solve many scientific and economic problems. For determining the humidity of the earth's crust [3]-[5], observing sea and ocean spaces [6, 7], researching climate changes of our planet [8, 9] and many others. In order to take into account the properties of the investigated surface when conducting practical experiments or modelling, it is necessary to know the relationship between the parameters of the investigated surface and the signals processed by the system. Moreover, the models of such connections are critically important when solving the inverse problem of estimating the parameters of an object or surface based on the signals registered by the radioengineering system.

II. Types of Surface Models

All radio engineering systems can be divided into two large classes: active and passive. Passive systems (radiometers) receive the thermal radiation of surfaces, while active systems (radars) irradiate the surface or object under study and receive the reflected signal, which already contains information about the parameters of this research object. According to the different nature of the appearance of electromagnetic waves, there are also different characteristics that connect these waves and surface parameters. In the passive case, this characteristic is the brightness temperature, in the active case, it is the complex reflection coefficient or radar cross section.

Also, all surface description models can be divided into electrodynamic (mathematical) [10]-[12] and empirical (practical) [13]-[19]. Mathematical models are well developed on the basis of the laws of electrodynamics and have been used for solving the problems of distance research for a certain period of time. Practical models are the result of specific experiments and are constantly being improved and developed.

For both electrodynamic and empirical models, the model has certain conditions and limitations for use, in particular, according to operating frequencies (wavelengths), type of cover (with or without vegetation) and geometric characteristics [20]: root mean square height of roughness or spatial height of roughness, radius of curvature of irregularities and the radius of correlation of irregularities.

In addition, there are models that allow taking into account the influence of the atmosphere on the propagation of emitted or reflected signals. And, as a separate task, the measurement of atmospheric parameters using remote sensing systems can be singled out.

A. Electrodynamic surface models

Electrodynamic models of the surface are not limited to any specific frequency range, but their use requires a very careful attitude to the relationship between the geometric

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

parameters of the investigated surface and the working wavelengths.

Figure 1 shows the dependences of the brightness temperature of a surfaces with large irregularities on the viewing angles for horizontal polarization for a system with a working wavelength of 0.03 m. The following surface parameters were used: thermodynamic temperature 300 (K), physical part of the dielectric constant – 4, medium conductivity – 0.1 (S/m), r.m.s. roughness height – 0.3 m [11, 21].

Figure 2 shows the dependences of the radar cross section of a surfaces with large irregularities with the same parameters on the viewing angles for horizontal polarization for a system with a working wavelength of 0.03 m.

Since electrodynamic descriptions for surfaces with different characteristics (a flat surface without irregularities, a finely rough surface, a surface with large irregularities, a two-scale surface) can be applied both to describe the brightness temperature and to specify the effective scattering surface, this allows the use of such models for research of surfaces by complex measurements of both active and passive remote sensing systems.

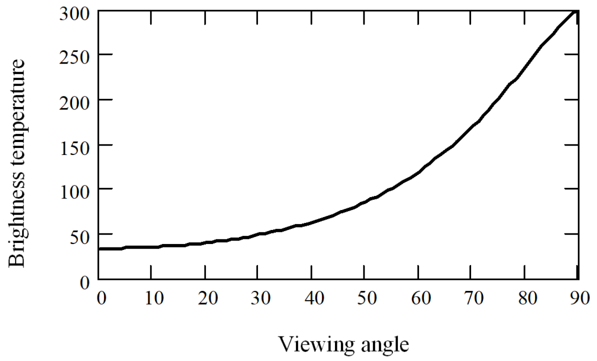


Fig. 1. Dependences of the brightness temperature on the viewing angle on the horizontal polarization.

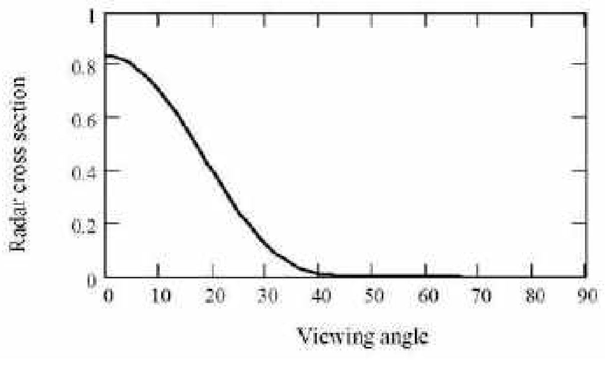


Fig. 2. Dependences of the radar cross section on the viewing angle on the horizontal polarization.

B. Empirical surface models

Table I shows the characteristics of some empirical models.

TABLE I. TYPES OF SURFACE MODELS

№	*Model name*	*Frequency range, GHz*	*Terms of use*
	Brightness temperature		
1	Model tau-omega	4-8,8	Vegetation is a uniformly absorbing and scattering layer above the soil surface
2	Sea surface model with foam	9,3-34	For a sea surface with foam (without taking into account atmospheric illumination) taking into account the wind speed
3	Qp-model	6,9-36,5	Surface without vegetation
4	Regression model	22,2-37,5	To estimate the moisture content of a cloudless atmosphere
	Effective scattering surface		
5	Exponential model	3-100	Quasi-smooth surfaces, rough without vegetation and with vegetation, as well as snow and anthropogenic areas
6	Surface model with vegetation	1-18	Surfaces with vegetation
7	Dubois model	1,5 - 11	Surface without vegetation, viewing angles from 30 to 65 degrees

Analytic expressions for models in Table I are taken from open sources and can be used to solve remote sensing problems. Of course, due to the large number of different types of land covers, such an analysis is very extensive, so in this paper only some models and only some their characteristics are given as an example [13]-[17].

In addition, there are models that allow taking into account the influence of the atmosphere on the propagation of emitted or reflected signals. And, as a separate task, the measurement of atmospheric parameters using remote sensing systems can be singled out.

Despite the fact that the empirical models were obtained as a result of specific experiments and have fairly clear limitations in use, the paper made it possible to describe the same surface using the empirical Qp-model and the Dubois model. They are presented in Table I and they have an overlapping operating frequency range and can be applied to surfaces without vegetation.

Figure 3 shows the dependences of the brightness temperature of surfaces described by the Qp-model on the viewing angles for horizontal (h) and vertical (v) polarization for a system with a working wavelength of 0.015 m.

Figure 4 shows the dependences of the radar cross section of a surface described by the Dubois model on the viewing angles for horizontal (h) and vertical (v) polarization for a system with a working wavelength of 0.015 m.

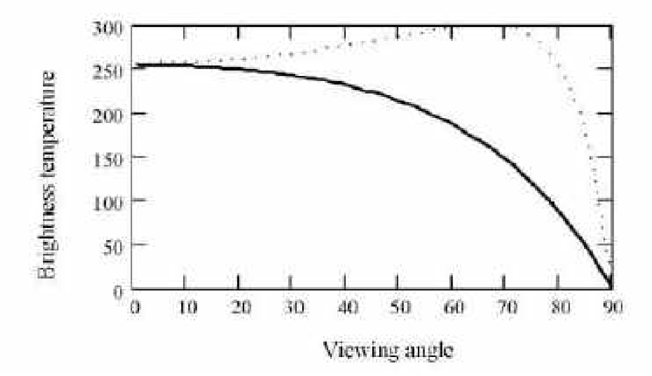


Fig. 3. Dependences of the brightness temperature on the viewing angle on the horizontal (solid line) and vertical polarizations (dot line).

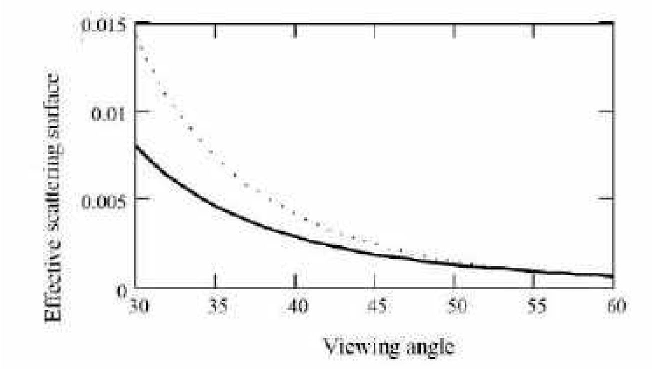


Fig. 4. Dependences of the effective scattering surface on the viewing angle on vertical (dot line) and horizontal polarization (solid line).

C. Classification of surface models

In this work, it is proposed to classify the currently known models from open sources according to the following classes:

- the type of radioengineering system where the model can be applied:
  - active type systems (the main characteristic of the surface is the effective scattering surface);
  - systems of the passive type (the parameter analysed is the antenna temperature and the brightness temperature associated with it);
- type of mathematical relations:
  - mathematical (electrodynamic, obtained on the basis of electrodynamic laws);
  - practical (empirical, regression, created as a result of practical experimental research);
- the type of physical state of the investigated surface:
  - ground surface:
    - surface with vegetation;
    - surface without vegetation;
    - snow covered surface and so on;
  - water surface:
    - fresh water;
    - salt water;
    - water with foam and so on.

Such a logical and consistent division into classes of various models for describing surfaces makes it possible to simplify and structure the process of selecting necessary expressions for passive and active remote sensing.

III. Algorithm

Based on the analysis and processing of a large amount of information [22]-[24] according to the proposed classification, an algorithm for selecting a surface model was created [10] (Fig. 5, Fig. 6) and a prototype of a software product for selecting the type of surface model was developed (Fig. 7). Thus, the user will be able to select functional dependencies for conducting certain experiments or calculations based on known initial data (type of system, operating frequencies, physical and geometric characteristics of the investigated surface).

At the first stage of the research, it is necessary to determine the given conditions: the researcher has information about the parameters of the system or about the type of the investigated surface. If the problem statement contains parameters and technical characteristics of the system, and it is necessary to determine its capabilities (variants of the investigated surfaces), the first algorithm is recommended for use (Fig. 5). If the task of the research is to study a specific type of surface (at least its geometric and physical characteristics are partially known), we use the second approach (Fig. 6), which allows us to choose options for the technical implementation of the research.

Next, it is necessary to choose the appropriate parameters of the system or surface, and as a result, recommendations for the application of the surface model will be obtained. Despite its simplicity, this algorithm will simplify and speed up some stages of remote sensing of the Earth's surface.

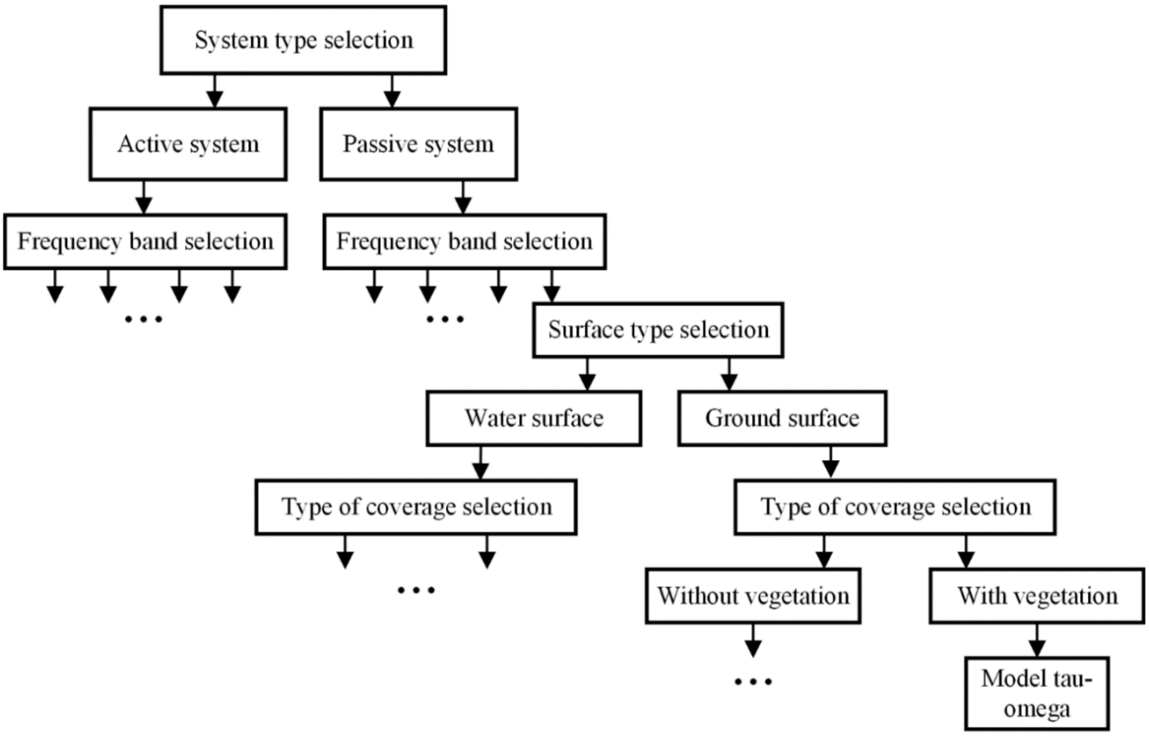


Fig. 5. Model selection algorithms when system type is stated.

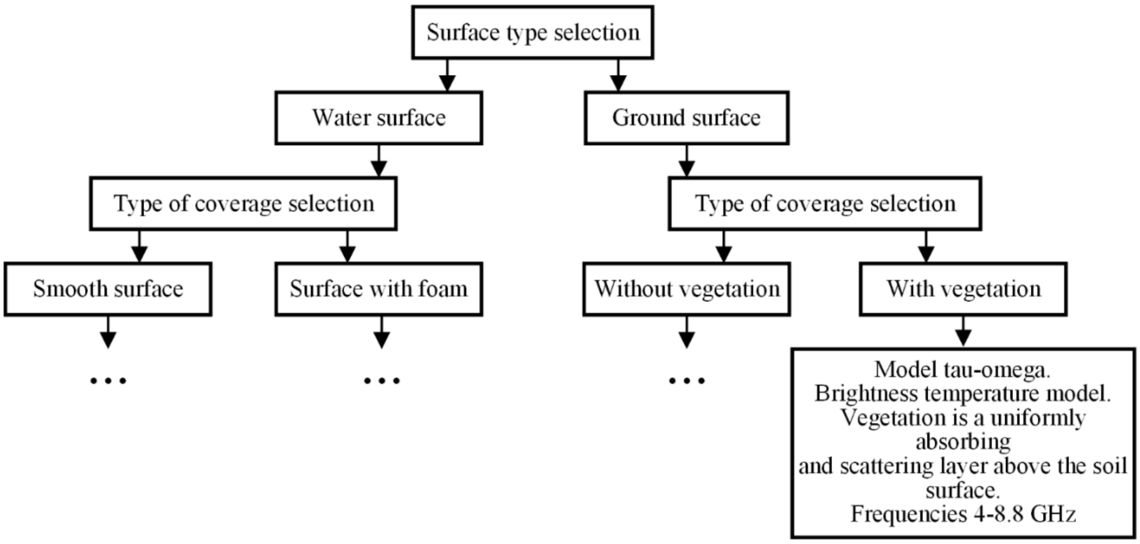

Surface type selection
Water surface
Ground surface
Type of coverage selection
Type of coverage selection
Smooth surface
Surface with foam
Without vegetation
With vegetation
...
...
...
Model tau-omega.
Brightness temperature model.
Vegetation is a uniformly absorbing
and scattering layer above the soil surface.
Frequencies 4-8.8 GHz

Fig. 6. Model selection algorithms when surface type is stated.

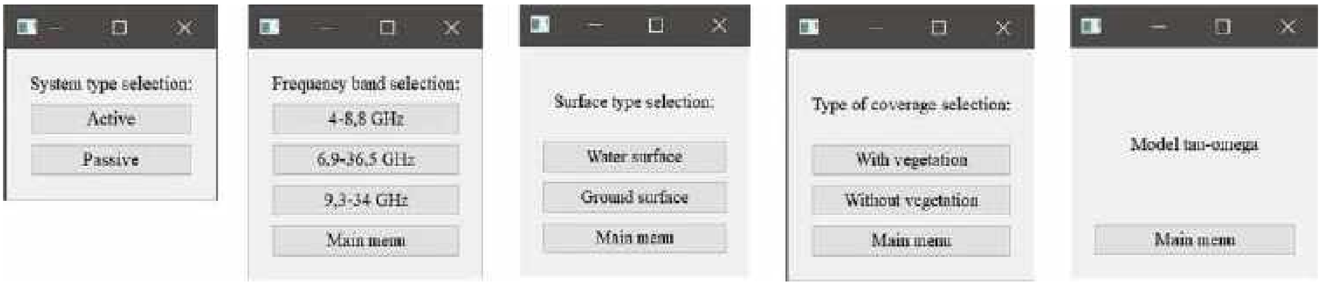


Fig. 7. Program interface fragment.

CONCLUSIONS

Thus, the presented work proposes a classification of models of surface parameters relationships with radiation or reflection signals while remote sensing by radio-technical means. An algorithm for simplifying the selection of such models and a software prototype for solving such a problem are proposed. Of course, both the practical and scientific side of remote research of various objects and surfaces is constantly developing and improving, therefore the proposed regulations and implementation are planned to be supplemented, clarified and refined.

REFERENCES

[1] S. Zhyla, V. Volosyuk, V. Pavlikov, N. Ruzhentsev, E. Tserne "Statistical synthesis of aerospace radars structure with optimal spatio-temporal signal processing, extended observation area and high spatial resolution," in Radioelectronic and computer systems, vol.101, issue 1, 2022, pp. 178-194.

[2] N. Ruzhentsev, S. Zhyla, V. Pavlikov, V. Volosyuk, E. Tserne, "Radio-Heat Contrasts of UAVs and Their Weather Variability at 12 GHz, 20 GHz, 34 GHz, and 94 GHz Frequencies," in ECTI Transactions on Electrical Engineering, Electronics, and Communications, vol 20, issue 2, 2022, pp. 163–173.

[3] Z. Hong, H. A. Moreno, Z. Li, S. Li, J. S. Greene, Y. Hong, L. V. Alvarez, "Triple Collocation of Ground-, Satellite- and Land Surface Model-Based Surface Soil Moisture Products in Oklahoma – Part I: Individual Product Assessment", Remote Sens., vol. 14 (22), 2022, 5641. https://doi.org/10.3390/rs14225641

[4] X. Wu, "Assessment of Effective Roughness Parameters for Simulating Sentinel-1A Observation and Retrieving Soil Moisture over Sparsely Vegetated Field", Remote Sens., vol. 14 (23), 2022, 6020. https://doi.org/10.3390/rs14236020

[5] V. Volosyuk, S. Zhyla, V. Pavlikov, N. Ruzhentsev, E. Tserne, "Optimal Method for Polarization Selection of Stationary Objects Against the Background of the Earth's Surface," in International Journal of Electronics and Telecommunications, vol 68, issue. 1, 2022, pp. 83-89.

[6] W. Sun, J. Wang, Y. Li, J. Meng, Y. Zhao, P. Wu, "New Gridded Product for the Total Columnar AtmosphericWater Vapor over Ocean Surface Constructed from Microwave Radiometer Satellite Data", Remote Sens., vol. 13 (12), 2021, 2402. https://doi.org/10.3390/rs13122402

[7] Wenqing Tang, Simon H. Yueh, Alexander J. Fore, Akiko K. Hayashi, Michael Steele, "An Empirical Algorithm for Mitigating the Sea Ice Effect in SMAP Radiometer for Sea Surface Salinity Retrieval in the Arctic Seas", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 11986-11997, 2021.

[8] Wilhelm May, "The role of land-surface interactions for surface climate in the ECEarth3 earth system model", Earth System Dynamics. Discussions. Preprint. Discussion started: 15 June 2023. https://doi.org/10.5194/esd-2023-13

[9] A. Popov, E. Tserne, V. Volosyuk, S. Zhyla, V. Pavlikov, "Invariant polarization signatures for recognition of hydrometeors by airborne weather radars," in. Computational Science and Its Applications – ICCSA 2023. Lecture Notes in Computer Science, vol.13956, 2023, Springer, Cham, pp. 1–14.

[10] V. K. Volosyuk, V. F. Kravchenko, "Statistical theory of radio engineering systems of remote sensing and radar (in Russian)", FIZMATLIT, Moscow, 2008, 704 p.

[11] U. A. Melnik, S.G. Zubkovich, V. D. Stepanenko, "Radar methods for exploring the Earth (in Russian)", Sov. Radio, 1980, 264 p.

[12] O. Shmatko, I. Ostroumov, N. Kuzmenko, K. Dergachov, O. Sushchenko "Synthesis of the optimal algorithm and structure of contactless optical device for estimating the parameters of statistically uneven surfaces," in Radioelectronic and computer systems, issue. 4, 2021, pp. 199-213.

[13] S. Paloscia,G. Macelloni, E. Santi, "Soil Moisture Estimates From AMSR-E Brightness Temperatures by Using a Dual-Frequency Algorithm", IEEE Transaction on Geoscience and Remote Sensing, vol. 44, №11, pp. 3135-3144, 2006.

[14] P.C. Pandly, R.K. Kakar, "An empirical microwave emissivity model for a foam covered sea", IEEE. I. of Oceanic Engineering, vol. OE-7, №3, pp. 135-140, 1982.

[15] Wei En-Bo, Ge Yong, "A microwave emissivity model of sea surface under wave breaking", Chinese Physics, vol. 14, №6, pp. 1259-1264, June 2005.

[16] J. Shi, L. Jiang, L. Zhang, K. S. Chen, J-P. Wigneron, A. Chanzy, T. J. Jackson, "Physically Based Estimation of Bare-Surface Soil Moisture With the Passive Radiometers", IEEE Transaction on Geoscience and Remote Sensing, vol. 44 (11), pp. 3145-3153, 2006. https://doi.org/10.1109/TGRS.2006.876706

[17] V. D. Stepanenko, G. G. Schukin, L. P. Bobylev, S. Yu. Matrosov, "Radiothermolocation in meteorology (in Russian)", Gidrometeoizdat, Leningrad, 1987.

[18] J. R. Wang, A. Hsu, J. C. Shi, P. E. O`Neill, E. T. Engman, "Estimating surface soil moisture from SIR-C measurements over the Little Washita River Watershed", vol. 59 (2), pp. 308-320, February 1997. https://doi.org/10.1016/S0034-4257(96)00145-9

[19] P. Dubois, J. Van Zyl, E.T. Engman E.T., "Measuring soil moisture with imaging radar", IEEE Trans. Geosci. Remote Sensing, vol. 3, №3, pp. 915-926, July 1995. http://dx.doi.org/10.1109/36.406677

[20] O. Solomentsev, M. Zaliskyi, Y. Averyanova, I. Ostroumov, N. Kuzmenko, "Method of Optimal Threshold Calculation in Case of Radio Equipment Maintenance," in Data Science and Security. Lecture Notes in Networks and Systems, vol. 462, 2022, Springer, Singapore, pp. 69–79.

[21] V. V. Bogorodsky, A. I. Kozlov, L. T. Tuchkov, "Radiothermal radiation of the earth's covers (in Russian)", Gidrometeoizdat, Leningrad, 1977, 226 p.

[22] O. Havrylenko K. Dergachov, V. Pavlikov, "Decision Support System Based on the ELECTRE Method," in Data Science and Security. Lecture Notes in Networks and Systems, vol. 462, 2022, Springer, Singapore, pp. 295–304.

[23] O. Sushchenko, Y. Averyanova, I. Ostroumov, "Algorithms for Design of Robust Stabilization Systems," in. Computational Science and Its Applications – ICCSA 2022. Lecture Notes in Computer Science, vol.13375, 2022, Springer, Cham, pp. 198–213.

[24] K. Dergachov, O. Havrylenko, V. Pavlikov, "GPS Usage Analysis for Angular Orientation Practical Tasks Solving," 2022 IEEE International Conference on Problems of Infocommunications. Science and Technology, Kyiv, Ukraine, 2022, pp. 1–6.

Analysis of Brightness Temperature Models for Describing Surfaces by Passive Remote Sensing Methods

Kseniia Nezhalska
Aerospace Radioelectronic Systems Department
National Aerospace University "Kharkiv Aviation Institute"
Kharkiv, Ukraine
k.nezhalska@khai.edu

Oleksandr Mazurenko
Aerospace Radio-Electronic Systems Department
National Aerospace University "Kharkiv Aviation Institute"
Kharkiv, Ukraine
o.mazurenko@khai.edu

Konstantin Belousov
Spacecraft, Measuring Systems and Telecommunications Department
Yuzhnoye SDO
Dnipro, Ukraine
https://orcid.org/0000-0002-6436-3359

***Abstract* — The National Aerospace University "Kharkiv Aviation Institute" has been conducting research for more than 10 years to optimize algorithms for processing radio-thermal radiation, improving the structures of microwave radiometers and developing own passive radars for aerospace applications. Last year it was developed ultra-wideband radiometer for a university nanosatellite with advanced bandwidth from 75 to 110 GHz. This passive radiometer is planned to be used to analyze the condition of the underlying Earth's surface. It is impossible to study the surface without a correct model of the relationship between its parameters, radiated electromagnetic fields and received signals. In this paper we analyze existing electrodynamic, empirical and regression models to describe the temperature of the underlying surface and determine the conditions of their application. The feature of the work is the study of the dependencies of the limiting errors of the developed ultra-wideband radiometer at different viewing angles, for different polarizations and for different estimated parameters.**

Keywords* — *remote sensing, electrodynamic surface models, empirical surface models, microwave radiometers

I. Introduction

The solution of the problems of monitoring the condition of the Earth's surfaces in modern science and practice, taking into account the constant climate change of our planet, is relevant and sometimes even necessary. Moreover, all modern studies, in particular studies of soil moisture [1, 2], use remote sensing methods and corresponding active [3] and passive radio engineering systems [4] to solve such problems.

Remote sensing in its broadest sense means obtaining information about the Earth's surface and objects on it or in its interior by means of various non-contact methods. At the same time, when organizing and using remote sensing data, the following tasks should be solved: models development for the mathematical relationship between surface parameters and scattering characteristics; compilation an observation equation; development of optimal processing algorithms; analysis of clear characteristics of estimates. Moreover, the choice of an optimal model for describing the surface and optimal conditions for conducting the experiment is the key to the efficiency of such an experiment and the interpretation of its results.

At the National Aerospace University "Kharkiv Aviation Institute" scientists have synthesised algorithms for optimal processing of radio-thermal signals and developed a microwave radiometer for remote surface sensing by passive method. It has unique characteristics in terms of bandwidth, which is 35 GHz and covers the range from 75 to 110 GHz. This radiometer will be a payload in the KhAI-1KA nanosatellite, which includes a platform developed at Yuzhnoye SDO. The structure of the microwave passive radar is shown in Fig. 1. It has the following components: 1 – W-Band Horn Antennas (WR-10, 75-110 GHz), 2 – 75-110 GHz PIN Diode modulator, 3 – WR-10 Low Noise Amplifier, 4 – 75 to 110 GHz RF Amplitude Detector.

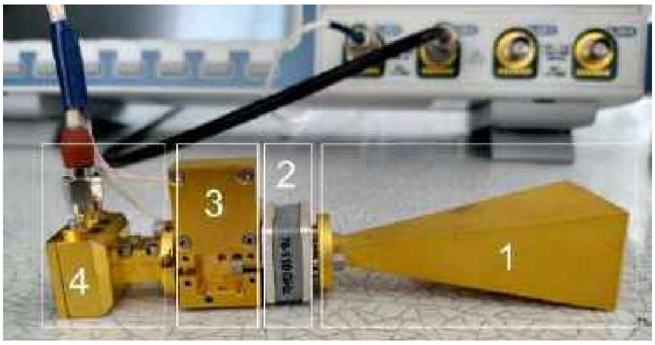


Fig. 1. Ultrawideband microwave radiometer for university nanosattelite

Having developed and analysed the technical characteristics of the radiometer, the team of authors faced the task of establishing a relationship between the measured values of the bright surface temperature using synthesized algorithms [5] and the characteristics of this surface. For this purpose, it is reasonable to analyse existing electrodynamic [6, 7], empirical [8, 9] and regression models [10] and determine the optimal sighting angles [11], polarisation [12, 13] and energy limits [14] to achieve the smallest marginal errors for measurements in the frequency range from 75 GHz to 110 GHz (wavelength 2.7 mm - 4 mm).

II. Analysis of Existing Models

From the mathematical and physical point of view, the relationship between the surface parameters and the radiation characteristics is most accurately described by the electrodynamic models [6, 7]. Such models can be applied to describe a water surface in complete calm (flat surface model), a surface of asphalt, concrete or arable land (small-scale roughness model), a desert or water surface with different degrees of excitement (large-scale and two-scale model) and other types of surfaces [15, 16]. Despite the electrodynamic

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

accuracy of such models, they do not always reflect all the properties of real Earth's surfaces, so the problem arose of creating more practical empirical or regression models. Empirical models are mainly created for the remote study of any particular type of Earth's surface or for conducting a particular experiment. Among such models we can mention, for example, the foam surface model [8, 9], model $\tau-\omega$ [17] and model Q_P [18].

In multi-parameter measurements in passive remote sensing systems, it is possible to estimate not only the underlying surface parameters, but also the atmospheric layer between this surface and the receiving antenna. For this purpose, the corresponding characteristics of this atmospheric layer must be introduced into the mathematical relationship between the surface parameters and the radiated field. Such studies can be performed using atmospheric flat surface models [6, 19] and regression models of moisture storage estimation [10].

For the mentioned electrodynamic and empirical models, a comparative analysis in terms of operating frequencies and limitations in use was performed and the results are summarised in Table 1. There is no frequency limitation on the application of electrodynamic models, but there are strict requirements for the ratio of the operating frequency (wavelength λ) and geometric parameters of the surface: the root mean square height of irregularities σ_h (or their spatial height $h(x,y)$), radii of curvature of irregularities R_κ, etc. Empirical models, as stated earlier, are created for specific experimental studies and work correctly only in limited frequency ranges and for a specific type of surface.

TABLE I. COMPARATIVE ANALYSIS OF MODELS

#	Model	Frequency f	Wavelength λ	Conditions of application
1	Flat	–	–	$h(x,y)=0$, $\sigma_h \approx 0$
2	Flat with atmosphere	–	–	$h(x,y)=0$, $\sigma_h \approx 0$
3	Small-scale	–	–	$\lvert h(x,y)\rvert << \lambda$, $\frac{\partial h(x,y)}{\partial x} << 1$, $\frac{\partial h(x,y)}{\partial y} << 1$, $\sigma_h \le \frac{\lambda}{20}$
4	Large-scale	–	–	$R_{\kappa x} >> \lambda$, $R_{\kappa y} >> \lambda$, $\sigma_h \ge \lambda$
5	Dual scale	–	–	$\sigma_{h1} \ge \lambda$, $\sigma_{h2} << \lambda$
6	Surface with foam	9,3-34 GHz	8,8 mm – 3,2 cm	For sea surface with foam (excluding backlighting) taking into account wind velocity
7	Model $\tau-\omega$	4-8,8 GHz	3,4-7,5 cm	Vegetation is an evenly absorbing and dispersing layer over the soil surface
8	Q_p model	6,9-36,5 GHz	8,3 mm – 4,3 cm	Surface without vegetation

Thus, the procedure for selecting a model to describe the surface under study is a rather complex and important operation that must be carried out before each practical experiment. In addition, knowledge of the relationship between the surface radiation and its parameters is necessary for solving the inverse problem of estimating the surface parameters from the received own radiation by radio system.

III. BRIGHTNESS TEMPERATURE MODELS

For radiometer operating frequencies of 75-110 GHz (wavelength 2.7-4 mm), a comparative analysis of the effect of the atmosphere on the bright surface temperature described by the flat model was performed.

The radio brightness temperature of thermal radiation of a flat surface is described by the expression

$$T_{Br(V,H)} = (1-\left|\dot{K}_{f(V,H)}\right|^2)T_0\,, \tag{1}$$

where $\dot{K}_{f(V,H)}$ are Fresnel coefficients,

$$\dot{K}_{fH} = \frac{\cos\theta-\sqrt{\dot{\varepsilon}-\sin^2\theta}}{\cos\theta+\sqrt{\dot{\varepsilon}-\sin^2\theta}},\ \dot{K}_{fV} = \frac{\dot{\varepsilon}\cos\theta-\sqrt{\dot{\varepsilon}-\sin^2\theta}}{\dot{\varepsilon}\cos\theta+\sqrt{\dot{\varepsilon}-\sin^2\theta}}, \tag{2}$$

where $\dot{\varepsilon}$ is a complex dielectric permittivity of the medium, $\dot{\varepsilon}=\varepsilon-j\cdot 60\cdot\lambda\cdot g$ (λ is a wavelength, g is a conductivity of the medium); θ is a probing angle; T_0 is a thermodynamic temperature of the surface.

The brightness temperature of a flat surface taking into account atmospheric illumination can be written as [6, 7]

$$\begin{aligned} T_{Br(V,H)} &= \chi_{(V,H)}(\dot{\varepsilon},\theta)K(h_0,\theta)T_0 + \\ &+\left|\dot{K}_{f(V,H)}(\dot{\varepsilon},\theta)\right|^2 K(h_0,\theta)T_{Br\,A}(\theta) + \\ &+T_A\left[1-K(h_0,\theta)\right], \end{aligned} \tag{3}$$

where $\chi_{(V,H)}(\dot{\varepsilon},\theta)=(1-\left|\dot{K}_{f(V,H)}(\varepsilon,\theta)\right|^2)$ is an emissivity of the surface, $T_{Br\,A}(\theta)$ is a radio brightness temperature of the total radiation of the atmosphere reflected from the surface in the direction θ, T_A is an average temperature of the atmosphere (about $30K$ lower than the temperature of the atmosphere on the Earth),

$$T_{Br\,A}(\theta) = T_A\left[1-e^{\frac{-1}{\cos\theta}(\chi_{ko}z_k+\chi_{\text{в}o}z_{\text{в}})}\right], \tag{4}$$

$$K(h_0,\theta) = \exp\left\{-\frac{1}{\cos\theta}(\chi_{ko}z_k+\chi_{\text{в}o}z_{\text{в}})\right\}, \tag{5}$$

where z_k, $z_{\text{в}}$ are characteristic heights of oxygen and water vapour absorption, ($z_k = 5{,}3$, $z_{\text{в}} = 2{,}1$ km), χ_{ko}, $\chi_{\text{в}o}$ are oxygen and water vapour absorption coefficients near the Earth's surface ($\chi_{ko} = 0{,}0018$, $\chi_{\text{в}o} = 0{,}002$).

Below in Fig. 2 and Fig. 3 it is shown the dependences of the bright temperature on the angle of view without taking into account the influence of the atmosphere at horizontal $Th(\theta)$, $Tha(\theta)$ and vertical $Tv(\theta)$, $Tva(\theta)$ polarisations.

Initial data are frequency 90 GHz, thermodynamic temperature 300 K. Surface type is water (relative permittivity $\varepsilon = 70$, conductivity $g = 5$ cm/m) [5].

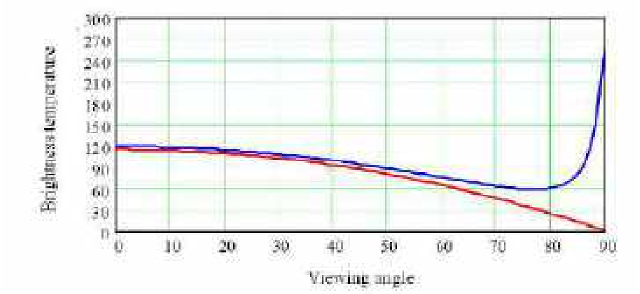


Fig. 2. Dependence of brightness temperature on the angle of view without (red line) and with (blue line) taking into account the influence of the atmosphere on horizontal polarisation.

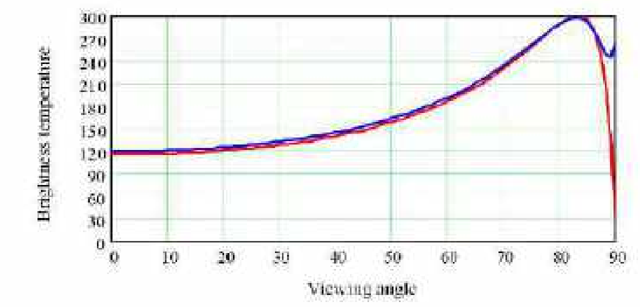


Fig. 3. Dependence of brightness temperature on the angle of view without (red line) and with (blue line) taking into account the influence of the atmosphere on vertical polarisation.

IV. Potential Accuracy Analysis

As mentioned earlier, one of the tasks that should be considered when organising an experiment and using remotely sensed data is the analysis of accurate estimation characteristics. Such an analysis can be performed using the Fisher information matrix [6, 19]. Such a study makes it possible to determine the conditions that ensure minimum errors in the measurement of surface parameters depending on polarisation, sighting angles and other parameters [20].

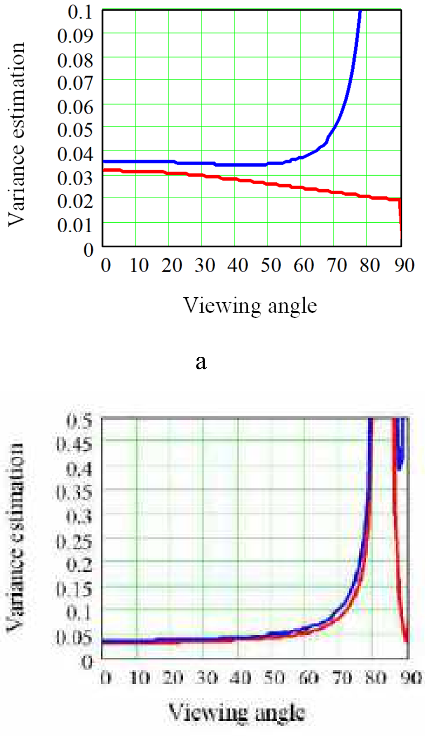
Let us consider the problem when it is necessary to estimate the limiting errors of measurements due to the electrophysical properties of the surface using a radiometer with a single antenna. In the simplest case, when one parameter is measured in one receiving channel, the marginal error dispersion (lower bound) is defined as follows

$$\sigma_{\lambda}^{2} = \frac{2}{T\Delta f}\left[\frac{T_{Br(V,H)}(\theta, f, \lambda)}{\partial T_{Br(V,H)}(\theta, f, \lambda)/\partial\lambda}\right]^{2}, \qquad (6)$$

where $\frac{2}{T\Delta f}$ is the parameter taking into account the observation time T and bandwidth Δf, f is the operating frequency, λ is the estimated parameter.

Below there are the graphs of dependence of the variance of estimation of the real component of the complex dielectric permittivity of the investigated surface on the angles of view θ [22, 23] without taking into account and with taking into account the influence of the atmosphere on the received radiation of horizontal $\sigma\varepsilon h(\theta)$, $\sigma\varepsilon ha(\theta)$ (Fig. 4a) and vertical $\sigma\varepsilon v(\theta)$, $\sigma\varepsilon va(\theta)$ (Fig. 4b) polarisations.

Calculation conditions: operating frequency 90 GHz, thermodynamic temperature 300^0 K, $\frac{2}{T\Delta f} = 10^{-6}$. Surface type: water (relative permittivity $\varepsilon = 70$, conductivity $g = 5$ cm/m).



b

Fig. 4. Variance estimation of the real component of the complex dielectric permittivity at horizontal (a) and vertical (b) polarisations (without (red line) and with (blue line) taking into account the influence of the atmosphere).

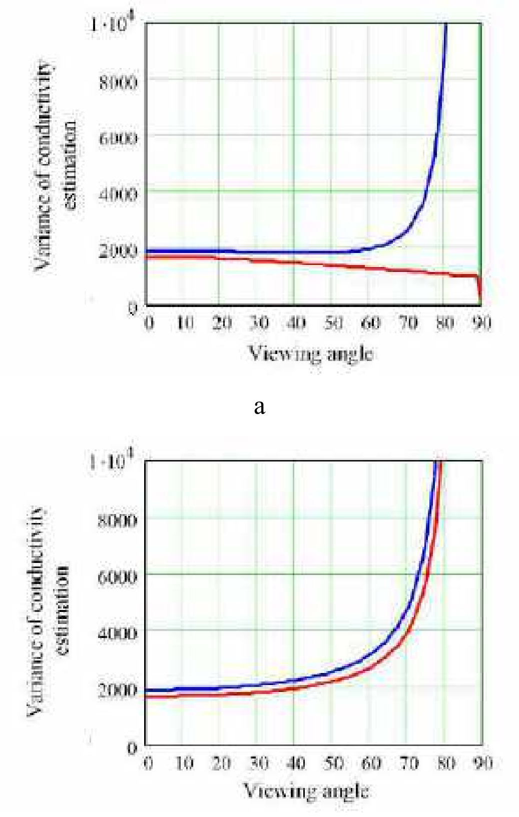
As can be seen from the obtained dependences, the accuracy of estimation of the real component of the complex dielectric permittivity ε is the highest at sighting angles θ from 0 to 60 degrees at both polarisations.

The estimated parameter can be the conductivity of the investigated surface g. Below plots of the dependence of the variance of the surface conductivity estimation on the sighting angles without and with taking into account the influence of the atmosphere on the received radiation of horizontal $\sigma gh(\theta)$, $\sigma gha(\theta)$ (Fig. 5a) and vertical $\sigma gv(\theta)$, $\sigma gva(\theta)$ (Fig. 5b) polarisations are given.

Calculation conditions: operating frequency 90 GHz, thermodynamic temperature 300^0 K, $\frac{2}{T\Delta f} = 10^{-6}$. Surface type: water (relative permittivity $\varepsilon = 70$, conductivity $g = 5$ cm/m).

The estimation of the surface permittivity does not give sufficient estimation accuracy in any of the given estimation methods. To solve this problem, it was proposed to estimate in general the imaginary part of the complex dielectric permittivity without singling out a separate conductivity g in it. In other words, to represent the complex dielectric permittivity in the form $\dot{\varepsilon} = \varepsilon - j \cdot 60 \cdot \lambda \cdot g = \varepsilon r - j \cdot \varepsilon i$. For this case, the potential accuracies of the estimation of the imaginary part of the complex dielectric permittivity without and with taking into account the atmosphere on horizontal $\sigma\varepsilon ih(\theta), \sigma\varepsilon iha(\theta)$ and vertical $\sigma\varepsilon iv(\theta), \sigma\varepsilon iva(\theta)$ polarisation are investigated (Fig. 6a, 6b).

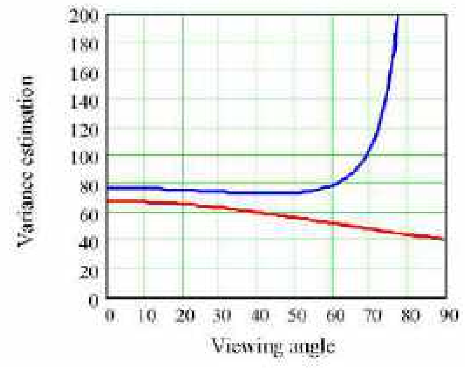
Calculation conditions: operating frequency 90 GHz, thermodynamic temperature 300^0 K, $\frac{2}{T\Delta f} = 10^{-6}$. Surface type: water (relative permittivity $\varepsilon = 70$, conductivity $g = 5$ cm/m, imaginary part $\varepsilon i = 1$).



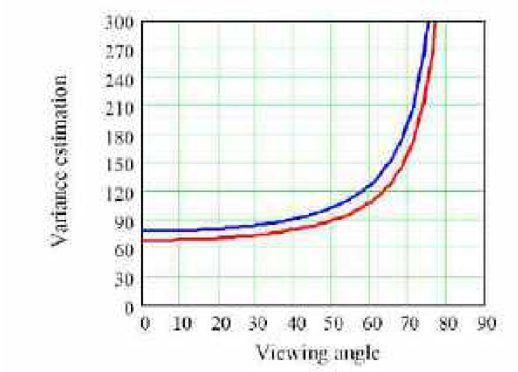
a

b

Fig. 5. Variance of conductivity estimation at horizontal (a) and vertical (b) polarisations (without (red line) and with (blue line) taking into account the influence of the atmosphere).



a



b

Fig. 6. Variance estimates of the imaginary component of the complex dielectric permittivity at horizontal (a) and vertical (b) polarisations (without (red line) and with (blue line) taking into account the influence of the atmosphere).

Obviously, the proposed approach provides a much better, but still insufficient, quality of assessment.

Conclusions

The presented research is performed for a radiometer with operating frequencies 75 - 110 GHz (wavelength 2.7 - 4 mm). The considered empirical bright temperature models work correctly at lower frequencies, and from the electrodynamic models the flat surface model is applicable. To fulfill the condition of using a small-scale surface ($\sigma_h \leq \frac{\lambda}{20}$) and at operating wavelengths of 2.7 - 4 mm, the root-mean-square height of surface irregularities should be less than 0.135 - 0.2 mm.

The influence of the atmosphere on the bright temperature, taken into account according to the proposed model, becomes evident at viewing angles greater than 50° in horizontal polarisation and angles greater than 80° in vertical polarisation.

The variance of conductivity estimation increases with increasing wavelength, while the variance of conductivity estimation decreases. For radiometers with the above mentioned operating frequencies it is possible to estimate the real dielectric permittivity or separately only the real component in the case of surface description by the complex permittivity parameter. Moreover, variance of real dielectric permittivity estimations taking into account the atmosphere on horizontal polarisation sharply increases at sighting angles greater than 60°. To estimate the surface conductivity (imaginary part of the complex dielectric permittivity of the surface), devices with lower operating frequencies are required. Moreover, the real dielectric permittivity of the surface can act as an estimated parameter, and it is recommended to investigate vertical and horizontal polarisations, without and taking into account the influence of the atmosphere.

References

[1] Z. Hong, H.A. Moreno, Z. Li, S. Li, J.S. Greene, Y. Hong, L.V. Alvarez, "Triple Collocation of Ground-, Satellite- and Land Surface Model-Based Surface Soil Moisture Products in Oklahoma—Part I: Individual Product Assessment," Remote Sens. 2022, 14, 5641. https://doi.org/10.3390/rs14225641

[2] X. Wu, Assessment of Effective Roughness Parameters for Simulating Sentinel-1A Observation and Retrieving Soil Moisture over Sparsely

Vegetated Field. Remote Sens. 2022, 14, 6020. https://doi.org/10.3390/rs14236020

[3] S. Zhyla, V. Volosyuk, V. Pavlikov, "Statistical synthesis of aerospace radars structure with optimal spatio-temporal signal processing, extended observation area and high spatial resolution," in Radioelectronic and computer systems, vol.101, issue 1, 2022, pp. 178-194.

[4] N. Ruzhentsev, S. Zhyla, V. Pavlikov, "Radio-Heat Contrasts of UAVs and Their Weather Variability at 12 GHz, 20 GHz, 34 GHz, and 94 GHz Frequencies," in ECTI Transactions on Electrical Engineering, Electronics, and Communications, vol 20, issue 2, 2022, pp. 163–173.

[5] S. Zhyla, V. Volosyuk, V. Pavlikov, N. Ruzhentsev, E. Tserne, "Practical imaging algorithms in ultra-wideband radar systems using active aperture synthesis and stochastic probing signals," in Radioelectronic and computer systems, vol.105, issue 1, 2023, pp. 55-73.

[6] U. A. Melnik, S.G. Zubkovich, V. D. Stepanenko, "Radar methods for exploring the Earth (in Russian)", Sov. Radio, 1980, 264 p.

[7] V. K. Volosyuk, V. F. Kravchenko, "Statistical theory of radio engineering systems of remote sensing and radar (in Russian)", FIZMATLIT, Moscow, 2008, 704 p.

[8] P.C. Pandly, R.K. Kakar An empirical microwave emissivity model for a foam covered sea // IEEE. I. of Oceanic Engineering. – 1982. – V. OE-7. – №3. – P. 135-140.

[9] Wei En-Bo, Ge Yong. A microwave emissivity model of sea surface under wave breaking // Chinese Physics. – 2005. – Vol 14. – №6, June. – P. 1259-1264.

[10] A.E. Basharinov, A.S. Gurvich, S.T. Egorov, Radioizluchenie Zemli kak planety (Microwave Emission of the Earth as a Planet), Moscow: Nauka, 1974, 187 p.

[11] O. Shmatko, I. Ostroumov, N. Kuzmenko, K. Dergachov, O. Sushchenko "Synthesis of the optimal algorithm and structure of contactless optical device for estimating the parameters of statistically uneven surfaces," in Radioelectronic and computer systems, issue. 4, 2021, pp. 199-213.

[12] A. Popov, E. Tserne, V. Volosyuk, S. Zhyla, V. Pavlikov, "Invariant polarization signatures for recognition of hydrometeors by airborne weather radars," in. Computational Science and Its Applications – ICCSA 2023. Lecture Notes in Computer Science, vol.13956, 2023, Springer, Cham, pp. 1–14.

[13] V. Volosyuk, S. Zhyla, V. Pavlikov, N. Ruzhentsev, E. Tserne, "Optimal Method for Polarization Selection of Stationary Objects Against the Background of the Earth's Surface," in International Journal of Electronics and Telecommunications, vol 68, issue. 1, 2022, pp. 83-89.

[14] O. Solomentsev, M. Zaliskyi, Y. Averyanova, I. Ostroumov, N. Kuzmenko, "Method of Optimal Threshold Calculation in Case of Radio Equipment Maintenance," in Data Science and Security. Lecture Notes in Networks and Systems, vol. 462, 2022, Springer, Singapore, pp. 69–79.

[15] V.K. Volosyuk, V.M. Velasco Herrera, K.N. Lyovkina, "Statistical relationship of mathematical models of the brightness temperature of radiothermal radiation and the reflection coefficient," Aviation and space technology and technology. - 2004. - No. 6 (14). - pp. 65-69.

[16] V.M. Velasco Herrera, G. Velasco Herrera, V.F. Kravchenko, V.K. Volosyuk, K.N. Lyovkina, "Radiothermal radiation of a small-scale surface. Investigation of the potential accuracy of measurements of its electrophysical parameters," Advances in modern radio electronics. Foreign radio electronics. - 2006. - No. 7. - pp. 60-69.

[17] Paloscia S., Macelloni G., Santi E. Soil Moisture Estimates From AMSR-E Brightness Temperatures by Using a Dual-Frequency Algorithm // IEEE Transaction on Geoscience and Remote Sensing. – 2006. – №11. – V. 44. – P. 3135-3144.

[18] Shi J., Jiang L., Zhang L., Chen K. S., Wigneron J-P., Chanzy A., Jackson T.J, Fellow Physically Based Estimation of Bare-Surface Soil Moisture With the Passive Radiometers // IEEE Transaction on Geoscience and Remote Sensing. – 2006. – №11. – V. 44. – P. 3145-3153.

[19] Volosyuk V.K. Theoretical foundations of passive remote sensing of natural environments from aerospace aircraf, KhAI Publishing House, 1997. - 84 p.

[20] Velasco Herrera G., Volosyuk V.K., Kurtov A.I., Lyovkina K.N. Investigation of radiothermal radiation of a small-scale surface and marginal errors in the estimation of its electrophysical parameters. Aviation and space engineering and technology. - 2005. - No. 5 (21). - pp. 70-78.

[21] O. Havrylenko K. Dergachov, V. Pavlikov, "Decision Support System Based on the ELECTRE Method," in Data Science and Security. Lecture Notes in Networks and Systems, vol. 462, 2022, Springer, Singapore, pp. 295–304.

[22] O. Sushchenko, Y. Averyanova, I. Ostroumov, "Algorithms for Design of Robust Stabilization Systems," in. Computational Science and Its Applications – ICCSA 2022. Lecture Notes in Computer Science, vol.13375, 2022, Springer, Cham, pp. 198–213.

[23] K. Dergachov, O. Havrylenko, V. Pavlikov, "GPS Usage Analysis for Angular Orientation Practical Tasks Solving," 2022 IEEE International Conference on Problems of Infocommunications. Science and Technology, Kyiv, Ukraine, 2022, pp. 1–6.

Application of Geoinformation Technologies for Modeling the Movement of Water in the River Network of the Selected Area

Yaryna Kokovska
Department of applied mathematics and informatics
Ivan Franko National University of Lviv,
Lviv, Ukraine
yaryna.kokovska@gmail.com

Mykola Prytula
Department of applied mathematics and informatics
Ivan Franko National University of Lviv,
Lviv, Ukraine
mykola.prytula@lnu.edu.ua

Mykhailo Oleksyn
Department of applied mathematics and informatics
Ivan Franko National University of Lviv,
Lviv, Ukraine
mykhailo.oleksyn@lnu.edu.ua

***Abstract* — The mathematical model of fluid movement in an open pseudo-prismatic channel is considered. A variational statement of the problem was formulated, which was solved by the finite element method. This problem is partially considered for the movement of water in the river basin. At each section of the river network, the channel characteristics may change, among which the angle of inclination of the middle bottom line plays an important role. The results are verified on test cases with complex bottom topography and show the influence of the choice of basis functions on the accuracy of the solutions and the calculated orders of convergence for temporal and spatial variables. Also, a GIS component was developed on the real part of the river basin, which allows showing the volume of water and the fullness of the channel in real time. By measuring the amount of precipitation in a given area, we can predict the fullness of the river bed, and this allows us to prevent critical water rises in the rivers of the selected basin and prevent such critical phenomena as floods, overflows of the river, inundation of adjacent territories. This contributes to the development of mountainous regions and the planning of production and industrial processes in them.**

***Keywords* — Navier-Stokes equations, finite elements methods (FEM), quadratic basis functions, bottom relief, river network, GIS component, basin river.**

I. Introduction

Channel runoff is one of the most important processes of the hydrological cycle, which includes rainwater and channel runoff, fluid runoff from the catchment surface, as well as a number of oceanological problems.

The formation of water flow from the surface of the catchment is a complex natural phenomenon, which is caused by a large number of factors [1,2,4,5,10,11]. Assessment and measurement of these factors are extremely difficult due to the presence of dependence on space and time. Therefore, the construction of a mathematical model of flow formation requires simplification and schematization of the main processes.

The most important problem for the practical application of the results of hydrodynamics is the problem of quickly issuing accurate forecasts of the hydrological situation in the studied section of the channel flow. It is for this that it is necessary to have a certain mathematical model of the river flow, which would accurately simulate the processes taking place in the river during a given period of time. After all, quickly obtaining the results of river flow modeling is an urgent task for emergency services in extreme situations. A large number of factors affecting the processes that take place in the river made it difficult to build a river model with the necessary parameters, and to obtain detailed information about the spatial distribution of at least the main characteristics of the catchment, which affect the inflow of water to the catchment, its infiltration and runoff.

In this paper a mathematical model of channel flow is described, for which an initial-boundary value problem and a variational formulation were constructed. For the created problem discretization was carried out according to spatial and temporal variables, evaluations of the convergence of the developed recurrent schemes were made, their stability was proved and verified on test examples.

II. Review of Existing Model runoff channel

Problems related to hydrological forecasts have certain specificities, which are as follows:

- solutions strongly depend on the future state of the weather, which is unknown at the time of the forecast release;
- operational information may be significantly less accurate and complete;
- fairly high accuracy of the method is required, since its effectiveness is evaluated relative to the inertial forecast;
- to obtain a significant practical result, it is necessary to deal with fairly large river catchments.

These features could not but affect the concepts underlying the models and the structure of the latter. In most cases, relatively simple models are used that are not very demanding on initial information.

Mathematical models of filling the reservoir with flows of inflowing rivers use relations based on one-dimensional Saint-Venant equations [2]. Analytical solutions to hydrodynamics problems are impossible to obtain in many practically important cases. Therefore, the following groups of methods are most often used to solve such problems [3]: finite-difference methods; small parameter methods; direct methods. These methods are implemented in MIKE11, MOSRIV packages [8,9]. FEM is also used to solve problems of hydrodynamics [12].

A reliable calculation of the characteristics of liquid flows in the river network can be performed in models that take into account the shape of the river bed and the presence of a constant inflow, as well as the turbulence of the flows at high speeds and irregularities at the bottom of the channels and changes in its trajectory.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

III. Equations of Water Flow in an Open Pseudoprizmatic Channel

In the case under consideration, we will limit ourselves to the analysis of flows of a viscous incompressible fluid along a plane that forms a dihedral angle *δ* with the horizontal plane of the earth. We will assume that the studied flow is generated, say, by intense rainfall or snowmelt.

The one-dimensional model describing the movement of liquid in an open pseudo-prismatic channel is described by the equations [1]:

$$\frac{\partial (UF)}{\partial x}+\frac{\partial F}{\partial t}=q;$$

$$\frac{1}{g}\frac{\partial U}{\partial t}+\frac{\alpha}{g}U\frac{\partial U}{\partial x}-\frac{\alpha-1}{g}\frac{U}{F}\frac{\partial F}{\partial t}+\frac{1}{B}\frac{\partial F}{\partial x}+\frac{U^2}{C^2R}=i, \qquad (1)$$

where the unknown values are *U* the flow velocity and *F* the cross-sectional area of the flow, *g* is the acceleration of gravity , C is the coefficient of Chezy, i = sin δ is the angle of inclination of the line of the middle bottom to the x axis; *B* is the width of the channel, *R* is the hydraulic radius, *α* is the parameter known as the average speed correction, *q*(*x; t*) is the side inflow.

System of equations has been supplemented by initial $U|_{t=0}=u_0(x)$, $F|_{t=0}=f_0(x)$ on [0, L] and boundary conditions *U (t, 0) = 0, F (t, 0) = 0* in this way, the initial-boundary problem of the unknown – flow speed *U* and cross-sectional area *F* was formulated.

IV. Numerical Solution of the Problem of Water Flow in an Open Channel

A. Constraction Of Variational Problems

Linear $\varphi\in V$ and quadratic $\psi\in V$ basis functions were used when constructing the variational, where space of allowable functions *V* was defined as $H:=L^2(\Omega),\ V:=\{v\in H^1(\Omega)\,|\,v(0)=0\}$.

Then variational formulation of system (1) was written as:

Asked: $u_0, f_0\in H$

Find a pair, such that: $(u,f)\in L^2(0,T;V\times V)$ *such, that*

$$\begin{cases} a(u,f,\varphi)+a(f,u,\varphi)+b(f',\varphi)=0; \\ \dfrac{1}{g}b(u',\psi)+\dfrac{\alpha}{g}a(u,u,\psi)+\dfrac{1}{B}c(f,\psi)+\dfrac{1}{C^2R}d(u,u,\psi)- \\ \qquad\qquad -\dfrac{\alpha-1}{g}d(w,f',\psi)=\langle l,\psi\rangle; \\ b(u(0)-u_0,\varphi)=0, b(f(0)-f_0,\psi)=0. \end{cases} \qquad (2)$$

where $a(u,f,\varphi)=\int\limits_{\Omega}u\frac{\partial f}{\partial x}\varphi dx;\ b(u,\varphi)=\int\limits_{\Omega}u\varphi dx;$

$c(u,\varphi)=\int\limits_{\Omega}\frac{\partial u}{\partial x}\varphi dx;\ d(u,f,\varphi)=\int\limits_{\Omega}uf\varphi dx$ are bilinear form and $l(\varphi)=\int\limits_{\Omega}i\varphi dx$ is the linear functional.

For solving variation problem (2) the finite element method was used [6,7]. The time and space variables discretization were made.

B. Discretization of the Variation Problem in time variable

The time interval [0,T] was divided into equal segments $\lfloor t_j, t_{j+1}\rfloor$ with a step $\Delta t=t_{j+1}-t_j,\quad j=0,...,N_T$. For approximation of solutions $u(x,t), f(x,t)\in L^2(0,T;V)$ such polynomials form was used

$$\begin{cases} u_{\Delta t}(x,t)=\{1-\omega(t)\}u^j(x)+\omega(t)u^{j+1}(x); \\ f_{\Delta t}(x,t)=\{1-\omega(t)\}f^j(x)+\omega(t)f^{j+1}(x); \\ t\in[t_{j,}t_{i+1}], j=0,1,...,N_T-1, \omega(t_j,t)=\dfrac{t-t_j}{\Delta t} \end{cases} \qquad (3)$$

where unknown functions are $u^j(x), f^j(x)\in V_h$.

For functional $l(x,t)\in V_h^1$ (2) next approximation was used

$$l_{\Delta t}(x,t)=l_{j+1/2}=l(t_{j+1/2},x). \qquad (4)$$

Then the problem (2) was rewritten as:

$$\begin{cases} b(f^{j+1/2},\varphi)+\Delta t\gamma[a(u^j,f^{j+1/2},\varphi)+a(u^{j+1/2},f^j,\varphi)+a(f^{j+1/2},u^j,\varphi)+a(f^j,u^{j+1/2},\varphi)]= \\ =-a(u^j,f^j,\varphi)-a(f^j,u^j,\varphi); \\ \dfrac{1}{g}b(u^{j+1/2},\psi)+\dfrac{\alpha}{g}\Delta t\beta[a(u^j,u^{j+1/2},\psi)+a(u^{j+1/2},u^j,\psi)]+\dfrac{1}{B}\Delta t\beta c(f^{j+1/2},\psi)+ \\ +\dfrac{2}{C^2R}\Delta t\beta d(u^j,u^{j+1/2},\psi)-\dfrac{\alpha-1}{g}d(w^j,f^{j+1/2},\psi)=\langle l_{j+1/2},\psi\rangle-\dfrac{\alpha}{g}a(u^j,u^j,\psi)- \\ -\dfrac{1}{B}c(f^j,\psi)-\dfrac{1}{C^2R}d(u^j,u^j,\psi), \end{cases} \qquad (5)$$

Were unknown values were denote as:

$$u^j=u^j(x), f^j=f^j(x);\ u^{j+1/2}=u^{j+1/2}(x)=\frac{u^{j+1}(x)-u^j(x)}{\Delta t};$$

$$f^{j+1/2}=f^{j+1/2}(x)=\frac{f^{j+1}(x)-f^j(x)}{\Delta t}.$$

Then recurrent scheme was written as

Given:

$\Delta t, \omega(t)=const>0,\ u^j, f^j\in V\times V.$

Find:

$u^{j+1}, f^{j+1}\in V\times V.$

such that:

$$\begin{cases} b(f^{j+1/2},\varphi)+\Delta t\gamma[a(u^j,f^{j+1/2},\varphi)+a(u^{j+1/2},f^j,\varphi)+a(f^{j+1/2},u^j,\varphi)+a(f^j,u^{j+1/2},\varphi)]= \\ =-a(u^j,f^j,\varphi)-a(f^j,u^j,\varphi); \\ \dfrac{1}{g}b(u^{j+1/2},\psi)+\dfrac{\alpha}{g}\Delta t\beta[a(u^j,u^{j+1/2},\psi)+a(u^{j+1/2},u^j,\psi)]+\dfrac{1}{B}\Delta t\beta c(f^{j+1/2},\psi)+ \\ +\dfrac{2}{C^2R}\Delta t\beta d(u^j,u^{j+1/2},\psi)-\dfrac{\alpha-1}{g}d(w^j,f^{j+1/2},\psi)=\langle l_{j+1/2},\psi\rangle-\dfrac{\alpha}{g}a(u^j,u^j,\psi)- \\ -\dfrac{1}{B}c(f^j,\psi)-\dfrac{1}{C^2R}d(u^j,u^j,\psi); \\ u^{j+1}=u^j+\Delta tu^{j+1/2}, f^{j+1}=f^j+\Delta tf^{j+1/2}. \end{cases} \qquad (6)$$

where $\gamma = \dfrac{(\omega^2,\xi)}{(\xi,1)}, \beta = \dfrac{(\omega^2,\eta)}{(\eta,1)}$,

$(\xi,1)\int\limits_{t_j}^{t_{j+1}}\xi(\tau)d\tau \neq 0, (\eta,1)\int\limits_{t_j}^{t_{j+1}}\eta(\tau)d\tau \neq 0$ are coefficients of a recurrent scheme.

C. Discretization of variation problem for spatial variables

The finite spaces approximations V_h of the space V with properties $\dim V_h \xrightarrow[h\to 0]{} \infty$ was chosen. After the semi discrete approximation of solution *(u,f)* was written as (u_h, v_h).

The basis functions $\{\varphi_i(x)\}_{i=1}^N$ and $\{\psi_i(x)\}_{i=1}^M$ of the space V_h were chosen as linear polynomials and quadratic functions accordingly.

The unknown values were given as decomposition through the basis functions and unknown coefficients:

$$u_h^j(x) = \sum_{i=1}^{M} U_i^j \psi_i(x), \quad f_h^j(x) = \sum_{i=1}^{N} F_i^j \varphi_i(x) \tag{7}$$

So, problem (6) was written as:

Given:

$$\Delta t, \gamma, \beta = const > 0;$$

$$u^j, f^j \in R^n.$$

Find:

$$u^{j+1}, f^{j+1} \in R^n,$$

such that:

$$\begin{pmatrix} B1+\Delta t\gamma A1\left(u^j\right)+\Delta t\gamma A2\left(u^j\right) & \Delta t\gamma A3\left(f^j\right)+\Delta t\gamma A4\left(f^j\right) \\ \frac{1}{B}\Delta t\beta C+\frac{\alpha-1}{g}D2(w^j) & \frac{1}{g}B2+\frac{\alpha}{g}\Delta t\beta\left(A5\left(u^j\right)+A6\left(u^j\right)\right)+\frac{1}{C^2R}2\Delta t\beta D1(u^j) \end{pmatrix}\begin{pmatrix} f^{j+1/2} \\ u^{j+1/2} \end{pmatrix}=$$

$$=\begin{pmatrix} -AP1\left(u^j,f^j\right)-AP2\left(f^j,u^j\right) \\ L_{j+1/2}-\frac{\alpha}{g}AP3(u^j,u^j)-\frac{1}{B}CP(f^j)-\frac{1}{C^2R}DP(u^j,u^j) \end{pmatrix};$$

$$u^{j+1}=u^j+\Delta t u^{j+1/2}, f^{j+1}=f^j+\Delta t f^{j+1/2}. \tag{8}$$

where

$$B1=\{b1_{ij}\}_{i,j}^N=\{b(\varphi_i,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L\varphi_i(x)\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$A1(u_h^k)=\{a1_{ij}(u_h^k)\}_{i,j=1}^N=\{a(u_h^k,\varphi_i,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L u_h^k(x)\frac{\partial(\varphi_i(x))}{\partial x}\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$A2(u_h^k)=\{a2_{ij}(u_h^k)\}_{i,j=1}^N=\{a(\varphi_i,u_h^k,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L\varphi_i(x)\frac{\partial(u_h^k(x))}{\partial x}\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$A3(u_h^k)=\{a3_{ij}(f_h^k)\}_{i,j=1}^N=\{a(\varphi_i,f_h^k,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L\varphi_i(x)\frac{\partial(f_h^k(x))}{\partial x}\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$A4(f_h^k)=\{a4_{ij}(f_h^k)\}_{i,j=1}^N=\{a(f_h^k,\varphi_i,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L f_h^k(x)\frac{\partial(\varphi_i(x))}{\partial x}\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$AP1(u_h^k,f_h^k)=\{ap1_i(u_h^k,f_h^k)\}_{i,j=1}^N=\{a(u_h^k,f_h^k,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L u_h^k(x)\frac{\partial(f_h^k(x))}{\partial x}\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$AP2(f_h^k,u_h^k)=\{ap2_i(f_h^k,u_h^k)\}_{i,j=1}^N=\{a(f_h^k,u_h^k,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L f_h^k(x)\frac{\partial(u_h^k(x))}{\partial x}\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$C=\{c_{ij}\}_{i,j}^N=\{c(\varphi_i,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L\frac{\partial\varphi_i(x)}{\partial x}\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$B2=\{b2_{ij}\}_{i,j}^N=\{b(\varphi_i,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L\varphi_i(x)\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$A5(u_h^k)=\{a5_{ij}(u_h^k)\}_{i,j=1}^N=\{a(u_h^k,\varphi_i,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L u_h^k(x)\frac{\partial(\varphi_i(x))}{\partial x}\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$A6(u_h^k)=\{a6_{ij}(u_h^k)\}_{i,j=1}^N=\{a(\varphi_i,u_h^k,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L\varphi_i(x)\frac{\partial(u_h^k(x))}{\partial x}\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$D1(u_h^k)=\{d_{ij}(u_h^k)\}_{i,j=1}^N=\{d(u_h^k,\varphi_i,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L u_h^k(x)\varphi_i(x)\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$D2(w_h^k)=\{d_{ij}(w_h^k)\}_{i,j=1}^N=\{d(w_h^k,\varphi_i,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L w_h^k(x)\varphi_i(x)\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$AP3(u_h^k,u_h^k)=\{ap3_i(u_h^k,u_h^k)\}_{i,j=1}^N=\{a(u_h^k,u_h^k,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L u_h^k(x)\frac{\partial(u_h^k(x))}{\partial x}\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$CP(f_h^k)=\{cp_i\}_{i,j}^N=\{c(f_h^k,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L\frac{\partial f_h^k(x)}{\partial x}\varphi_j(x)dx\right\}_{i,j=1}^N;$$

$$DP(u_h^k,u_h^k)=\{dp_i(u_h^k,u_h^k)\}_{i,j=1}^N=\{d(u_h^k,u_h^k,\varphi_j)\}_{i,j=1}^N=\left\{\int\limits_0^L u_h^k(x)u_h^k(x)\varphi_j(x)dx\right\}_{i,j=1}^N.$$

V. Analysis of Approximations for Solving Problems

Since the problem is presented in the form of a nonlinear system of equations, the solution becomes oscillating at large angles of inclination of the bottom line. In such cases, it is advisable to increase the order of approximation of the unknown solution.

Example. The relief of the bottom is shown in Fig. 1.

Input data: α=1, $0 \le x \le 1$, $0 \le t \le 2$, $\Delta t = 0.007$, B=8, g=9.8, C=60, R=1

The results obtained when solving the system of equations using linear approximations are shown in Fig. 2-5. They show the influence of complex relief on the solution in the form of oscillations.

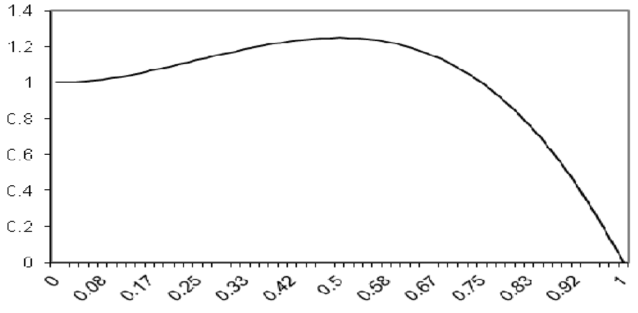


Fig. 1 Bottom surface of flow

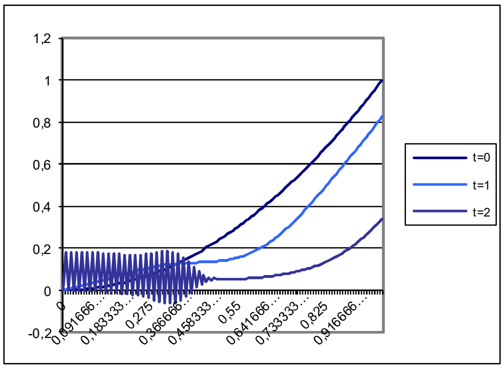


Fig. 2 Cross-sectional area of flow (linear approximation)

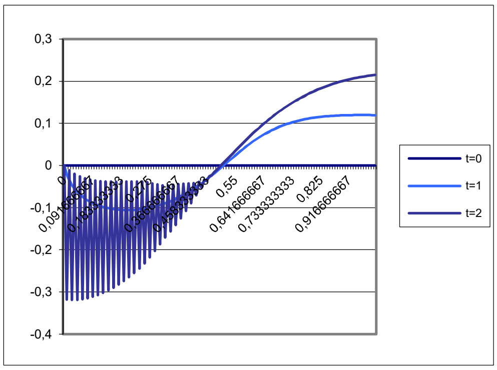


Fig. 3 Speed of flow (linear approximation)

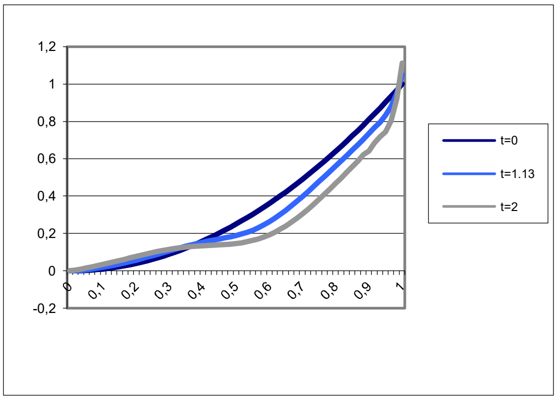


Fig. 4 Cross-sectional area of flow (quadratic approximation).

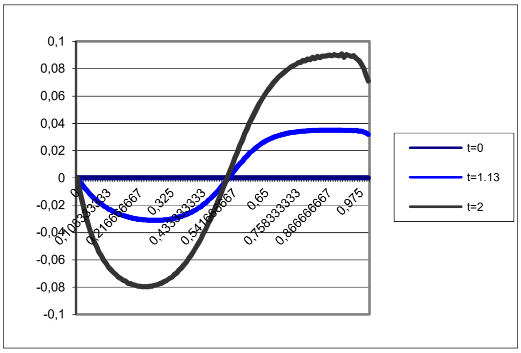


Fig. 5 Speed of flow (quadratic approximation)

The dynamics of changes in river beds, their depth, and flow velocity for equations (1) are shown in the graphs above.

This example with different angles of inclination of the line of the middle bottom shows the problems that we encounter when solving nonlinear equations (1) using linear approximation (Fig. 2-3) and the need to increase them to the second order, after which the graphs of changes in such parameters, such as the cross-sectional area of the flow and the velocity obtained in fig. 4-5.

VI. Calculation of Channel Flow for the River Network

The behavior of the water flow in the river network and the influence of changes in the trajectory of the riverbeds and the connections of the riverbeds with different characteristics were studied. These conditions were tested on test examples. In the case of water movement in a curved channel the division of the channel into straight parts was used (Figs. 6, 7), for which a local coordinate system with the main axis OX directed in the direction of the flow is introduced. A solution to the system of equations (1) was sought on each straight segment.

The transition from one segment to another is carried out by rotating the coordinate system

$$(x,y,1)\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ -x_0 & -y_0 & 1\end{pmatrix}\begin{pmatrix}\cos\varphi & \sin\varphi & 0\\ -\sin\varphi & \cos\varphi & 0\\ 0 & 0 & 1\end{pmatrix}=$$

$$=\left((x-x_0)\cos\varphi-(y-y_0)\sin\varphi,\ (x-x_0)\sin\varphi-(y-y_0)\cos\varphi,\ 1\right)$$

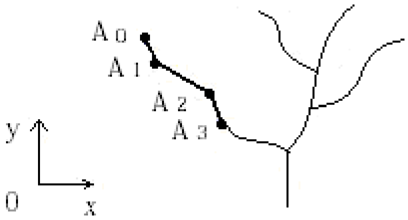


Fig. 6 Structure of river channel

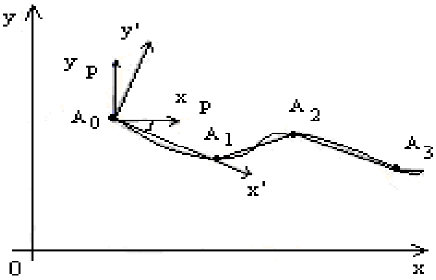


Fig. 7 Changing the local system coordinates

We show conditions for network connections rectilinear parts of the river.

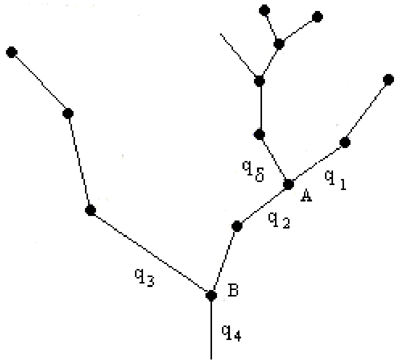


Fig. 8 Scheme of river network

Conditions of channel flows connections

in point A $q_2 = q_1 + q_\delta$, $h_1 = h_2$; q_δ lateral tributary

in point B $q_4 = q_3 + q_2$, $h_4 = h_2 = h_3$.

Using the conditions specified above, the movement of water in the river network and the conditions for its replenishment from a side tributary, as well as when connecting channels with different characteristics, were investigated and tested on examples.

Fig. 9 Electronic map of the Pivdennyi Bug river.

VII. TEST EXAMPLE

Input results: the length of flow x: 1; number of breaks per x: 20; length by t: 1; number of breaks by t: 50;

g = 9,8; B = 20; R = 1; C = 60.

Initial conditions: $u_0(x) = x$, $f_0(x) = x$, $\sin\delta = 0.02$.

The riverbed of the Bug River (Fig. 9) was chosen to obtain the solutions.

Examples with different sines of the angle of inclination of the line of the middle bottom will be considered. The following results were obtained:

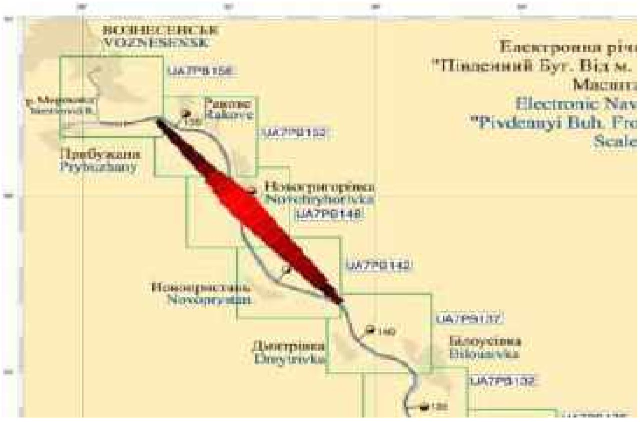


Fig. 10 Change in the cross-sectional area of the river current in the channel with a variable angle of inclination of the middle line of the bottom.

As a result of the calculations the following results were obtained Fig. 11.

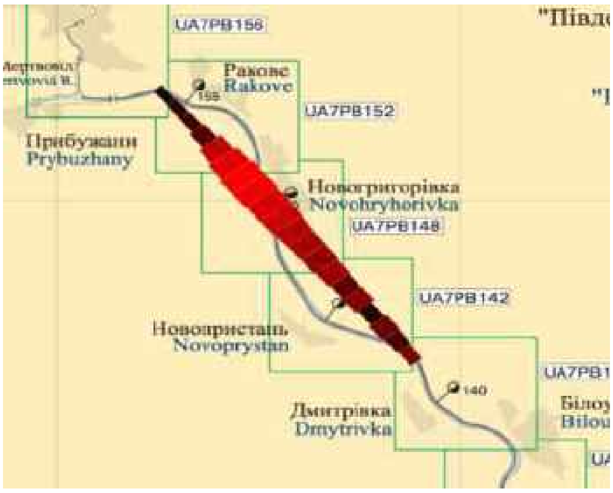


Fig. 11 Change of the speed of the river current in the channel with a variable angle of inclination of the middle line of the bottom.

Another example, with a constant sine of the angle of inclination along the entire channel, was also considered.

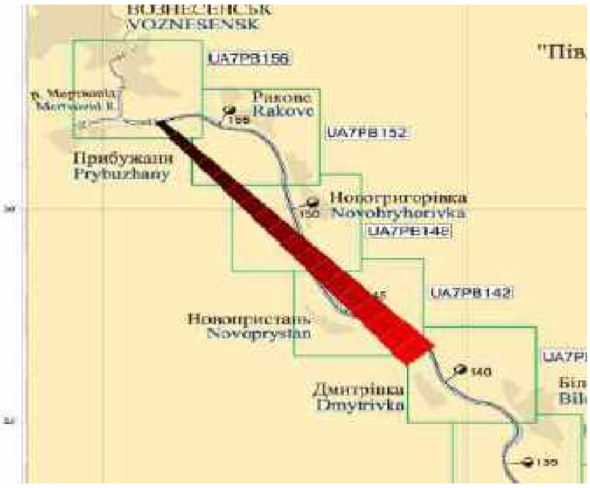


Fig. 12 Change of the cross-sectional area of the river flow with the angle of incline sin δ = 0.5.

So, the examples showed the behavior of water in riverbeds with a constant and variable value of the angle of inclination of the middle bottom, the flow of fluid in the river network was investigated and the results of numerical modeling of liquid movement in open channels were analyzed.

CONCLUSIONS

This paper describes a model of water movement in an open channel, which is described by a system of equations with unknown velocity and cross-sectional area of the flow. The finite element method was used to solve it. The movement of fluid in a river network with various side tributaries is considered. The analysis of the results was carried out on test cases with different input data. A GIS component for modeling the movement of water in the river basin was also built and its application on real maps was shown.

REFERENCES

[1] Subhasish Dey. Fluvial Hydrodynamics. Hydrodynamic and Sediment Transport Phenomena. Springer-Verlag, Berlin, Heidelberg, 2015. 687 p.

[2] Zh. Zhang, X. Song, Sh. Ye, Yi. Wang, “Application of deep learning method to Reynolds stress models of channel flow based on reduced-order modeling of DNS data,” Journal of Hydrodynamics, 2019, 31(1). pp. 58-65.

[3] L.S. Kuchment, Models of processes for the formation of river flow L.: Gidrometeoizdat, 1980. 142 p.

[4] M.S. Grushevsky Unsteady movement of water in rivers and channels. - L .: Gidrometeoizdat, 1982. - 288 p.

[5] M.A. Kartvelishvili Unsteady open flows. - L .: Gidrometeoizdat, 1968. - 126 p.

[6] Y. Kokovska, M. Prytula, P. Venherskyi, “Application of finite elements method for solving variational prolems of channel flows,” Journal of numerical and applied mathematics, 2017, Vol. 3(126), pp. 75– 85.

[7] Y. Kokovska, P. Venherskyi, “Investigation of the stability for established flows in open pseudoprismatic channels,” Eureka: physics and engineering. Computer sciences and mathematics, 2016, Vol. 5, pp. 9–15.

[8] I. T. Selezov, V. V. Kuznetsov, D. O. Chernikov, “Generation of surface gravity waves by bottom timerepetitive pulses,” J. Math. Sci. 2010. 171, N 5. P. 596–602.

[9] Y.F. Yao, T.G. Thomas, N.D. Sandham, “Direct Numerical Simulation of Turbulent Flow over a Rectangular Trailing Edge,” Theoret. Comput. Fluid Dynamics. 2001. No. 14. P.337– 358.

[10] P.S. Venherskyi, Ya.V. Kokovska, “One of the approaches to modeling the processes of channel flow of fluid,” Visn. Lviv univ. Applied Math. Inform. Num. 15 2009. P.178-195.

[11] O.Z. Zienkiewicz, R.L. Taylor, J.M. Too, “Reduced integration technique in general analysis of plates and shells,” Int. J. Num. Meth. Eng. 1971. – Vol. 3. P.275– 290.

[12] Ya.G. Savula Numerical analysis of problems of mathematical physics by variational methods. - Lviv .: LNU Publishing Center. I. Franko, 2004. - 221 p.

Noise-immune Transfer of Decimal Data with Protection Based on Permutations

Oleksiy A. Borysenko
Sumy State University
st. Rimsky-Korsakov, 2
Sumy, Ukraine
5352008@ukr.net

Oleksii Y. Horiachev
Sumy State University
st. Rimsky-Korsakov, 2
Sumy, Ukraine
a.goriachev@ekt.sumdu.edu.ua

Olga V. Berezhna
Sumy State University
st. Rimsky-Korsakov, 2
Sumy, Ukraine
o.berezhna@ekt.sumdu.edu.ua

Svitlana M. Matsenko
Sumy State University
st. Rimsky-Korsakov, 2
Sumy, Ukraine
s.matsenko@ekt.sumdu.edu.ua

Anatolii I. Novhorodtsev
Sumy State University
st. Rimsky-Korsakov, 2
Sumy, Ukraine
a.novhorodtsev@ekt.sumdu.edu.ua

Abstract — **The article proposes a method that solves the problem of hiding decimal numbers from unauthorized access while simultaneously detecting error packets in them and correcting single errors. The method is based on the encoding of decimal numbers by binary-coded permutations. With this encoding, each digit of a decimal number is first converted into a binary-coded permutation using a special encryption table and, after further mixing with other permutations, is transmitted to the receiver. On the receiving side, it is checked for errors and, after correcting them with the help of a key, it is converted into decimal digits. Since each digit of a decimal number contains 1 of 10 numbers, their encoding requires 10 permutations and, accordingly, at least 4 permutation elements: 0, 1, 2, 3. These elements form set of 24 permutations, which consists of 10 used and 14 redundant permutations. This redundancy, as well as the natural redundancy of binary-coded permutations, allows them to detect packets of errors and correct single errors.**

Keywords — Information protection, errors, noise immunity, numerical codes, secrecy, permutations

I. Introduction

BCD codes, known as binary coded decimal codes, are commonly used to obtain and transmit information from various sensors. These sensors provide data on various measurements, such as water temperature and electricity consumption. Typically, each BCD digit obtained from a sensor is transmitted through a communication channel, which can be either wired or mobile using radio communication [1]. In the case of mobile communication, data can be directly transmitted to moving objects like cars.

To ensure the security of transmitted information, the BCD digits of each decimal place are mixed using specific tables. These tables, essentially the initial keys of the cipher, allow the recipient to reconstruct the original information. Furthermore, the secrecy of each binary-decimal digit can be enhanced by additional mixing of the digits themselves.

Apart from protecting against unauthorized access, it is often necessary to enhance the resistance to noise for transmitted BCD numbers. BCD coding provides some degree of protection against interference due to its redundancy. However, the level of protection against interference is relatively low, although it may be acceptable for certain practical cases. Therefore, there is a need to improve the noise immunity of the system.

To address this issue, in [1-4] it is proposed to use binary-decimal error-correcting codes for the transmitted digits. Each decimal digit is encoded with an error-correcting combination, increasing the system's ability to detect errors [1-4]. This method introduces equilibrium code combinations for coding binary-decimal digits, which also enhances information security. Unlike textual information, which relies on statistical probabilities of letters for decoding, the statistical properties of equilibrium code combinations provide little assistance.

However, eliminating errors in the transmission of decimal digits using equilibrium code combinations is challenging, especially in mobile communications where retransmission can be difficult. Therefore, the practical goal is to develop a telecommunications system that not only detects errors but also corrects them. Additionally, this system should work with inseparable codes to conceal the actual value of decimal digits during transmission.

II. Problem Statement

The aim of this study is to enhance the resistance to noise of transmitted binary-decimal digits, while incorporating error correction, and ensure adequate protection against unauthorized access. To achieve this, it is proposed to increase the noise immunity of binary-decimal information using inseparable codes based on permutations. These codes serve a dual purpose: on the one hand they enable error detection and correction, and on the other hand they conceal the true information more securely.

Permutations are extensively utilized in mathematics, specifically in abstract algebra and solving combinatorial optimization problems like the traveling salesman problem [5,6]. Their applications continue to expand. Beyond their mathematical utility, permutations are successfully employed in practical information security challenges to protect data against unauthorized access [7-12].

Furthermore, permutations prove to be effective in error-correcting coding, as they inherently contain redundant information. This property facilitates the detection and elimination of errors in messages, which is particularly crucial for small mobile devices [13,14]. Moreover, permutations offer a means to combine solutions for error-correcting coding problems with effective information protection against unauthorized access.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

III. System for transmission and display of binary-decimal information

The structure of the system for transmitting and displaying binary-decimal information, for which noise-correcting coding based on permutations can be used, is shown in fig. 1 [1]. This circuit demonstrates the processing of a single BCD digit. It consists of both data transmission units and data reception units. On the transmitting side, the system comprises a control system (CU), a buffer memory circuit (BMS), an encoder (E). On the receiving side are an error correction block (ECB), a code converter (CC) of four-bit code combinations to seven-bit code combinations, and an indicator (I). The transmitting and receiving sides are connected by a communication line (CL).

The system works as follows. The bits of the BCD digit X1, X2, X3, X4 are received and stored at the BMS input. A four-digit binary-decimal digit from the output of the BMS is fed unchanged to the input of the encoder. The word received at the output of the encoder, which, as a rule, does not correspond to the incoming digit, is fed to the communication line CL, where it can be distorted by interference. If the garbled four-digit word does not relate to the words specified in the cipher table in its numerical value, then it is defined in the ECB as prohibited. Accordingly, this block issues a signal to the CU control circuit that an error has occurred. From the control circuit, a signal is sent to the BMS, and the input digit is resent to the CL.

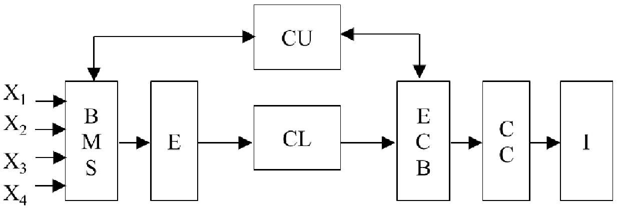


Fig. 1. BCD digit transmission and display system

After the ECB has accepted the signal as correct, it sends it to the CC input, which converts the code to the form intended for display on indicator I. The digit displayed on the indicator must correspond to the BCD digit stored in the BMS. The supply of symbols for indication to the BMS occurs at a given frequency, that is, they change periodically.

The transmission and display of BCD digits that form digits in a decimal number can occur in parallel or sequentially [1]. Parallel transmission and display on indicators requires a separate communication channel to transmit the contents of each bit, which increases the hardware costs of the transmission and display system as a whole. Sequential transmission of BCD digits of a multi-digit number one after another with their display on one indicator makes the system cheaper but increases the transmission time of the digits. The display time, on the contrary, decreases, which creates inconvenience for the operator.

IV. Coding of information by permutations

A permutation is any finite sequence of n distinct elements. At the same time, any symbols can be elements of permutations, but numbers are most often used as them. For example, a sequence of four distinct numbers 0 1 2 3 would be a permutation of length n = 4.

The set consisting of n! permutations of length n, forms a code on permutations. Difference $n \bullet \log_2 n - \log_2 n!$ forms redundant information for the permutations of this code, which can reach a significant value with increasing n, which determines the high noise immunity of the permutations. In addition, an important property of permutations is that their elements do not repeat and, therefore, obtaining their statistics is difficult. As a result, permutations can effectively protect the information contained in them from unauthorized access.

When solving problems of error-correcting coding and information protection, the elements of permutations are represented in binary form. Such a representation of them will be called binary-coded. Encoding BCD information would require 10 different binary-coded permutations. Therefore, the minimum value of n that can provide the required number of permutations is 4, since 4! = 24 > 10. Of the 24 permutations for encoding BCD numbers, 10 permutations are used, each of which encodes one of these numbers. The remaining 14 possible permutations are redundant. One of the possible representations of BCD numbers by permutations is shown in Table 1. Together, the BCD numbers in this table form a BCD code (2-10 code), and the corresponding binary-coded permutations form a binary permutation code with a permutation length of 8 bits.

TABLE I. Coding by permutations

№	2–10 code	Permutation	Binary permutation code
1	0000	0123	00 01 10 11
2	0001	0132	00 01 11 10
3	0010	0213	00 10 01 11
4	0011	0231	00 10 11 01
5	0100	0312	00 11 01 10
6	0101	0321	00 11 10 01
7	0110	1023	01 00 10 11
8	0111	1032	01 00 11 10
9	1000	1203	01 10 00 11
10	1001	1230	01 10 11 00

V. Information secrecy

Coding information by permutations provides protection against unauthorized access due to the large number of variants of such a code [15]. The number of options for choosing 10 permutations for encoding binary-decimal numbers is equal to the number of combinations of 10 out of 24. Each of these options in turn can be distributed among the encoded numbers 10! ways. One of the possible options for encoding of binary-decimal numbers by permutations is given in Table 1. To increase the secrecy of the transmitted information, the digits of decimal numbers, the number of which is equal to k, can also be rearranged in various ways. Accordingly, the total number of options for encrypting a decimal code based on permutations will be equal to

$$M = k! \bullet 10! \bullet C^{10}_{24} \quad (1)$$

The dependence of the number of permutation options M, containing from 1 to 20 digits of the binary-decimal number in the transmitted decimal number k, is shown in Table 2. It determines the complexity of the cipher, which grows exponentially with the growth of M. The key to the cipher is the number of the permutation in bits, the width of the key is equal to the logarithm of M. It should be considered that the statistics of digits in the cipher on permutations is poorly expressed, which significantly complicates its disclosure outside the enumeration of keys.

TABLE II. NUMBER OF PERMUTATIONS M

k	M	Key length in bits	k	M	Key length in bits
1	$7{,}11 \cdot 10^{12}$	43	11	$2{,}84 \cdot 10^{20}$	68
2	$1{,}42 \cdot 10^{13}$	44	12	$3{,}40 \cdot 10^{21}$	72
3	$4{,}27 \cdot 10^{13}$	46	13	$4{,}43 \cdot 10^{22}$	76
4	$1{,}70 \cdot 10^{14}$	48	14	$6{,}20 \cdot 10^{23}$	80
5	$8{,}54 \cdot 10^{14}$	50	15	$9{,}30 \cdot 10^{24}$	83
6	$5{,}12 \cdot 10^{15}$	53	16	$1{,}48 \cdot 10^{26}$	87
7	$3{,}58 \cdot 10^{16}$	55	17	$2{,}53 \cdot 10^{27}$	92
8	$2{,}86 \cdot 10^{17}$	58	18	$4{,}55 \cdot 10^{28}$	96
9	$2{,}58 \cdot 10^{18}$	62	19	$8{,}65 \cdot 10^{29}$	100
10	$2{,}58 \cdot 10^{19}$	65	20	$1{,}73 \cdot 10^{31}$	104

VI. EVALUATION OF NOISE IMMUNITY OF BINARY-DECIMAL CODE ON PERMUTATIONS

Permutations, in addition to providing secrecy, offer an improvement in the noise immunity of the BCD code. The binary representation of such permutations contains four elements, each consisting of m=2 bits, as shown in Table 1. Then the total length in bits of the binary-coded permutations will be 8. Each binary-coded permutation differs from the others by at least two bits, thereby establishing a minimum code distance of 2. This code distance enables the detection of any single error and all errors of odd multiplicity, such as 1, 3, 5, and so on.

The noise immunity of the code based on binary-coded permutations can be estimated by evaluating the fraction of detected errors (D) [13]. D represents the probability of mistakenly transforming a binary-coded permutation into a forbidden combination that is known to be detectable. It is determined by dividing the number of forbidden combinations (Z_f), which is 246, by the total number of combinations (Z), which is equal to 256. Accordingly, for binary-coded permutations $D = 246 / 256 = 0.96$.

VII. ERROR DETECTION ALGORITHM

A transmission error can translate a binary-coded permutation either into a forbidden combination that is not a permutation, or into one of the forbidden permutations. In the case where an error transforms a permutation into a forbidden combination that is not a permutation, it can be detected since the sum of the elements of the permutation must remain constant, equal to

$$S = n \bullet (n - 1) / 2. \qquad (2)$$

It can be used to identify erroneous combinations for which the sum of elements does not match the value determined by formula (2) [13, 14]. For the considered code on permutations S = 4 • (4 - 1) / 2 = 6.

Example 1. On the receiving side, when transmitting a binary-coded permutation, after converting it into permutation elements, their sequence 1231 was received. Calculating the sum of these elements gives the result 1 + 2 + 3 + 1 = 7. This number does not coincide with the checksum value obtained above for a code on permutations S = 6. This means that the resulting sequence is not a permutation and contains an error.

VIII. ERRORS CORRECTION METHODS BASED ON MODULO 2 ADDITION OF PERMUTATIONS

In [14], a method is considered that allows error detection and correction in one of the permutation elements. Its operation is based on the properties of permutations, without requiring additional coding. To do this, from all 24 permutations of length n = 4, a set of 10 allowed permutations is selected, which differ from each other by three elements. This ensures that minimum code distance between their binary representations is equal to 4, which satisfies the condition for correcting single errors. Thus, the error can be corrected by algorithm specified in [14].

Another method for correcting errors in permutations, which can be applied to any set of permutations. The method works when several digits of a decimal number are transmitted as an array of binary-coded permutations. Together with this array of permutations, their modulo 2 checksum is transmitted. On the receiving side, each permutation is checked for errors. For this, the calculation of the arithmetic sum of the elements of the permutation is used. If the receiving side detects an error in one of the permutations, this error is corrected as follows. The modulo 2 sum of all error-free binary-coded permutations and the checksum is calculated. The result will be the correct value of the permutation received with the error [15]. In the case when there is more than one error in the transferred permutations, such errors can only be detected using the method under consideration, but not corrected.

Example 2. On the transmitting side, five digits of the decimal number are encoded with the permutations shown in Table 3. Their modulo 2 sum is calculated, indicated in Table 3 as XOR.

TABLE III. TRANSMITTED INFORMATION

Permutation	Binary permutation code
0123	00 01 10 11
0231	00 10 11 01
0312	00 11 01 10
1203	01 10 00 11
1320	01 11 10 00
XOR	00 01 10 11

At the receiving side, the received information is checked for errors, as shown in Table 4. For each permutation, the sum of the elements of S is calculated. The correct value of the sum of the elements of any permutation of 4 elements, as indicated above, is 6. The result obtained shows that an error occurred in the second permutation, where S = 8 ≠ 6.

TABLE IV. CORRECTION OF ERROR IN PERMUTATION

Permutation	S	Binary permutation code
0123	6	00 01 10 11
0233	8	00 10 11 11
0312	6	00 11 01 10
1203	6	01 10 00 11
1320	6	01 11 10 00

Since only one of the permutations contains an error, such an error can be corrected. To do this, the XOR checksum and error-free binary-coded permutations are modulo 2 added. As a result, the correct value of the erroneous permutation is obtained, as shown in table 5.

TABLE V. CHECKING RECEIVED INFORMATION FOR ERRORS

Permutation	Binary permutation code
0123	00 01 10 11
0312	00 11 01 10
1203	01 10 00 11
1320	01 11 10 00
XOR	00 01 10 11
Result	00 10 11 01

Thus, the proposed method for correcting errors in code on permutations contains the following steps:

1. In the received array of binary-coded permutations, the arithmetic sum of all elements of each permutation is calculated. If the value of the sum does not match the value obtained by formula (2), the permutation contains an error.

2. If only one of the received binary-coded permutations contains an error, such an error can be corrected. Otherwise, retransmission of erroneous permutations is required.

3. To correct single error in a binary-coded permutation, the modulo 2 sum of XOR checksum and all permutations from the received array, except for the erroneous one, is calculated. The result of this calculation replaces the erroneous permutation.

The proposed error correction method can be used for any permutation coding table. In addition, it allows to correct not only single errors, but also all errors of odd multiplicity, provided that only one permutation contains an error.

CONCLUSIONS

The paper introduces an inseparable permutation code as a means of encoding decimal digits, which addresses the challenge of covertly transmitting digital information through a communication channel while maintaining noise immunity. The information's secrecy in the given scenarios achieves acceptable values, as permutations can conceal the statistical characteristics of transmitted decimal digits.

In addition to ensuring secrecy, permutations offer an effective solution for enhancing the noise immunity of transmitted digits. They facilitate the detection and correction of error packets. Furthermore, the methods employed for error detection and correction in permutations are relatively straightforward to implement in hardware, which is an important practical consideration.

REFERENCES

[1] O. Borisenko, O. Berezhna, A. Novgorodtsev, V. Serdiuk, M. Yakovlev Information transmission and display system with protection of numerical data. Information Processing Systems. 2019. № 2(157). pp. 103-108.

[2] O. Borysenko, V. Kalashnikov, Chapter 7: "Description and applications of binomial numeral systems complex" in Security and noise immunity of telecommunication systems: new solutions to the codes and signals design problem: Collective monograph. ASC Academic Publishing, Minden, Nevada, 2017, pp. 147-159.

[3] A. Kuznetsov, R. Serhiienko, D. Prokopovych-Tkachenko, B. Akhmetov, "Chapter 3: Representation of cascade codes in the frequency domain" in Security and noise immunity of telecommunication systems: new solutions to the codes and signals design problem: Collective monograph. ASC Academic Publishing, Minden, Nevada, 2017, pp. 71-101.

[4] A. Kuznetsov, S. Ksvun, Y. Gorbenko, "Chapter 4: The methodology of evaluating the energy gains from coding in channels with grouping errors" in Security and noise immunity of telecommunication systems: new solutions to the codes and signals design problem: Collective monograph. ASC Academic Publishing, Minden, Nevada, USA, 2017, pp. 102-119

[5] D.Knuth, The Art of Computer Programming, Vol. 1: Fundamental Algorithms, 3rd ed., Addison-Wesley Professional, 1997.

[6] D.Knuth, The Art of Computer Programming, Vol. 4A: Combinatorial Algorithms, Part 1, 1st ed., Addison-Wesley Professional, 2011.

[7] D. Smith, R. Montemanni, "A new table of permutation code", Designs, Codes and Cryptography, Vol. 63, pp. 241-253, 2012.

[8] W. Stallings, Cryptography and Network Security Principles and Practices, fourth ed., Prentice Hall, 2005.

[9] R. Girija, H. Singh, "A new substitution-permutation network cipher using Walsh Hadamard Transform" in Proceedings of International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017, pp. 168-172. DOI: 10.1109/IC3TSN.2017.8284470

[10] A. Aryal, S. Imaizumi, T. Horiuchi, H.i Kiya, "Integrated algorithm for block-permutation-based encryption with reversible data hiding" in Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017, pp. 203 - 208. DOI: 10.1109/APSIPA.2017.8282028

[11] I. Janiszczak, R. Staszewski, "An improved bound for permutation arrays of length 10", [On-line]. Available: http://www.iem.uni-ue.de/preprints/IJRS.pdf [October 16, 2014].

[12] J. Barta, R. Montemanni, "Hamming Graphs and Permutation Codes" in Proceedings of Fourth International Conference on Mathematics and Computers in Sciences and in Industry (MCSI), 2017, pp. 154 - 158. DOI: 10.1109/MCSI.2017.35

[13] O. Borysenko, O. Horiachev, S.Matsenko, O.Kobiakov, "Noise-immune codes based on permutations" in Proceedings of 9th International IEEE Conference «Dependable Systems, Services and Technologies DESSERT'2018», 2018, pp. 645-648. DOI: 10.1109/DESSERT.2018.8409204

[14] O. Borysenko, O. Horiachev, V. Serdyuk, A. Horyshnyak, O. Kobyakov, O. Berezhna, "Protection of numerical information based on permutations" in Proceedings of International Scientific And Practical Conference "Information Security And Information Technologies". Kharkiv – Odesa: Simon Kuznets Kharkiv National University of Economics, 2021, pp. 68-73.

[15] O. Borysenko, O. Horiachev Methods of error detection and correction in permutations. Information Processing Systems. 2013. – №2 (109). – pp. 171-173.

Analysis of the Periodically Non-stationary Structure for Modulated Vibration Signal

Ihor Javorskyj
Department of Methods and Facilities
for Acquisition and Processing Diagnostic Signals
Karpenko Physico-mechanical Institute of NAS of Ukraine,
Lviv, Ukraine
ihor.yavorskyy@gmail.com

Roman Yuzefovych
Department of Methods and Facilities for Acquisition and
Processing Diagnostic Signals
Karpenko Physico-mechanical Institute of NAS of Ukraine,
Lviv, Ukraine
Department of Applied Mathematics,
Lviv Polytechnic National University, Lviv, Ukraine
roman.yuzefovych@gmail.com

Oleh Lychak
Department of Methods and Facilities
for Acquisition and Processing
Diagnostic Signals
Karpenko Physico-mechanical Institute
of NAS of Ukraine,
Lviv, Ukraine
oleh.lychak2003@yahoo.com

Pavlo Semenov
Department of Hoisting and Transport
Machines and Engineering of Port
Technological Equipment
Odesa National Maritime University
Odesa, Ukraine
p.a.semenoff@gmail.com

Mykola Varyvoda
Department of Methods and Facilities
for Acquisition and Processing
Diagnostic Signals
Karpenko Physico-mechanical Institute
of NAS of Ukraine,
Lviv, Ukraine
mykola.zen.varyvoda@gmail.com

***Abstract* — The vibration time series are analyzed using methods for periodically non-stationary random processes (PNRP). Band-pass filtering and Hilbert transform are used to extract quadrature components. The cross-covariance structure of quadratures is considered. It was shown that cross-covariances of different order quadratures result in periodical non-stationarity of the vibration signal.**

Keywords* — *periodically non-stationary random signals, vibrations, Hilbert transform, quadratures, modulation.

I. Introduction

The analysis of vibrations is widely used for testings of the rotating machinery [1–11]. The vibration signals which are generated by damaged mechanism can be described as a set of stochastically modulated carrier harmonics [1–4, 10, 11]. The different signal processing methods are used for analysis of modulation properties [12–23]. Widely used so-called "envelope detection" methods in vibration signal diagnostic, which involves Hilbert transform are heuristically introduced without strong mathematical basis are more "state of the art" than strong measurement methodology [10-12, 17, 18, 24, 25]. Using PNRP [14] we can analyze in more detail the covariance and spectral structures of stochastic modulations of PNRP carrier harmonics. Separation of the individual modulated harmonics and their quadratures can be performed using band-pass filtering with following Hilbert transform of results [24]. Properties of the Hilbert transform of vide-band vibration signals in most known works are analyzed only superficially [25, 26]. Involvement of the PNRP harmonic series representation to Hilbert transform usage of the simplest PNRP particular case allows obtain much more promising results.

In this work using PNRP techniques we analyze the characteristic feature of signal selected from damaged mechanism.

II. Vibration signal of a damaged rotary unit

The vibration (acceleration) signal of a port decanter bearing unit was acquired and pre-processed using the original vibro-diagnostic system. Analog signal was low-pass filtered to 5 *kHz*. Sampling rate 10 *kHz*. The acquired series length was *T=10 s*. A fragment of acquired series is depicted in Fig. 1. Step-by-step algorithm for processing a high-frequency modulated time series described by a PNRP is represented in a flowchart in Fig. 2.

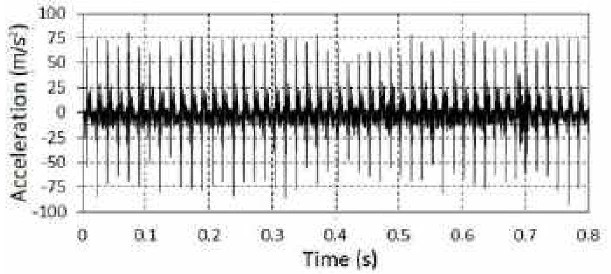


Fig. 1. Fragment of the acquired vibration signal

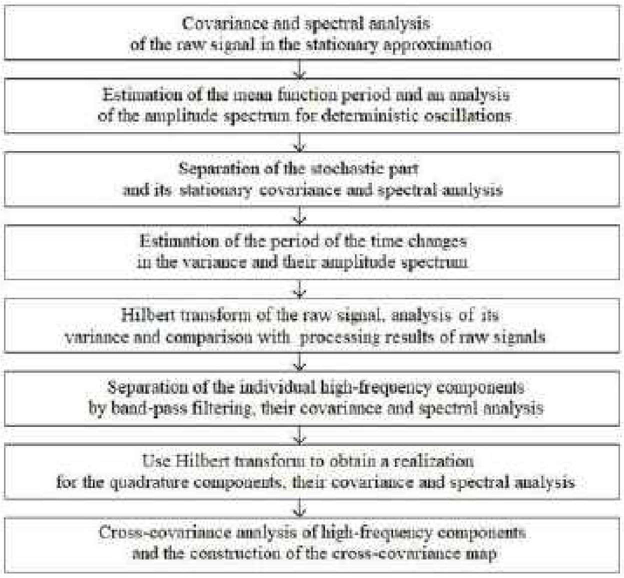

Covariance and spectral analysis
of the raw signal in the stationary approximation
Estimation of the mean function period and an analysis
of the amplitude spectrum for deterministic oscillations
Separation of the stochastic part
and its stationary covariance and spectral analysis
Estimation of the period of the time changes
in the variance and their amplitude spectrum
Hilbert transform of the raw signal, analysis of its
variance and comparison with processing results of raw signals
Separation of the individual high-frequency components
by band-pass filtering, their covariance and spectral analysis
Use Hilbert transform to obtain a realization
for the quadrature components, their covariance and spectral analysis
Cross-covariance analysis of high-frequency components
and the construction of the cross-covariance map

Fig. 2. Flowchart for procedure of PNRP analysis

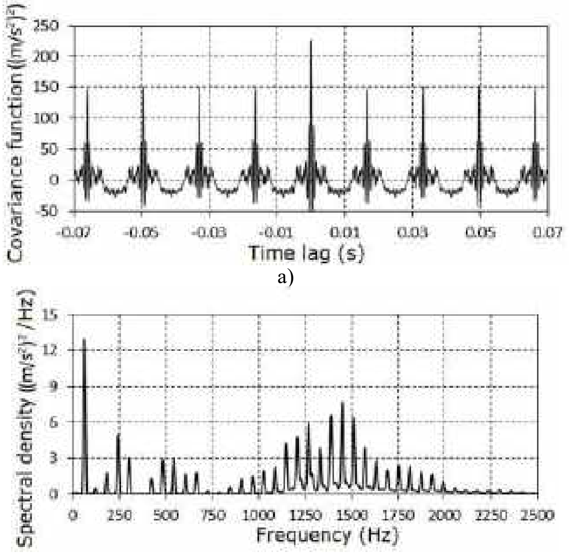
To study the properties of obtained time series $\xi(nh)$, the covariance function estimator and the power spectral density estimator for the stationary approximation were obtained using following equations:

$$\hat{R}(jh)=\frac{1}{K}\sum_{n=0}^{K-1}\left[\xi(nh)-\hat{m}\right]\left[\xi((n+j)h)-\hat{m}\right],$$
$$\hat{m}=\frac{1}{K}\sum_{n=0}^{K-1}\xi(nh), \qquad (1)$$

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

$$\hat{f}(\omega)=\frac{h}{2\pi}\sum_{n=-L}^{L}k(nh)\hat{R}(nh)\cos\omega nh\,. \qquad (2)$$

Here, $h=T/K$ is the sampling interval, j is the integer number, K is number of samples, $L=u_m/h$, u_m is the cut-off point of the correlogram. $k(nh)$ term denotes a Hamming window. Estimator of the covariance function and the spectral density estimator in (1) and (2) depicted in Fig. 3.



a)

b)

Fig. 3. The covariance function estimator (a) and the power spectral density estimator (b) in stationary approximation

The covariance estimator contains of the undamped tail, which represents the discrete component of the spectral density estimator. It is depicted by the peaks in certain frequencies. These peaks could be a result of the narrow-band stochastic modulation of the carrier harmonics.

III. Analysis of Signal Mean Function

The mixed spectrum nature leads to difficulties in interpretation of the spectral density estimation and its quantitative analysis. To estimate the period θ of the mean function that describes the deterministic oscillation, we use the functional [15, 27–30]:

$$\hat{F}_1(\theta)=\frac{1}{2K+1}\sum_{n=-K}^{K}\hat{m}^2(\theta,nh), \qquad (3)$$

where

$$\hat{m}(\theta,nh)=\sum_{k=1}^{L_1}\left[\hat{m}_k^c(\theta)\cos k\frac{2\pi}{\theta}nh+\hat{m}_k^s(\theta)\sin k\frac{2\pi}{\theta}nh\right], \quad (4)$$

$$\begin{Bmatrix}\hat{m}_k^c(\theta)\\ \hat{m}_k^s(\theta)\end{Bmatrix}=\frac{2}{2K+1}\sum_{n=-K}^{K}\xi(nh)\begin{Bmatrix}\cos k\dfrac{2\pi}{\theta}nh\\ \sin k\dfrac{2\pi}{\theta}nh\end{Bmatrix}, \qquad (5)$$

and L_1 is the number of chosen harmonics. The graph of the dependency of the functional in (3) on the test frequency $f=1/\theta$ depicted in Fig. 4. The maximum point on Fig. 4 corresponds to the basic frequency; its estimation is 60,430 *Hz*. Using estimated value of this frequency, we calculated the quantities in (4, 5) and the mean function (Fig. 5) using interpolation formula:

$$\hat{m}_{\xi}\left(t,\hat{P}\right)=\hat{m}_0+\sum_{k=1}^{L_1}\left[\hat{m}_k^c\left(\hat{P}\right)\cos k\frac{2\pi}{\hat{P}}t+\hat{m}_k^s\left(\hat{P}\right)\sin k\frac{2\pi}{\hat{P}}t\right].$$

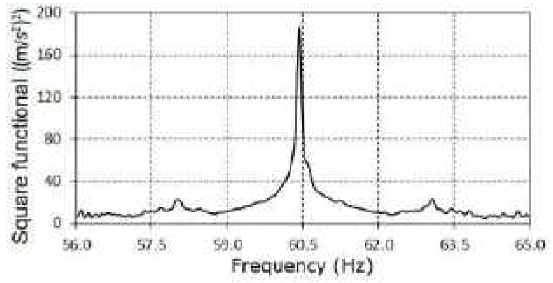
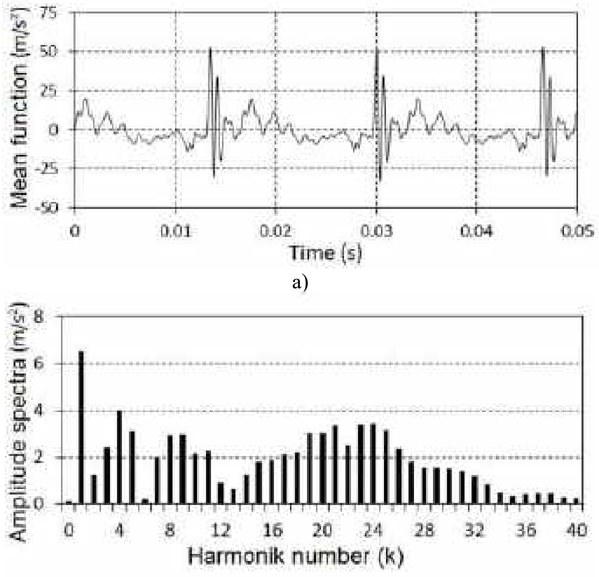


Fig. 4. Dependency of calculated functional (3) on test frequency



a)

b)

Fig. 5. Estimator of the mean function (a) and its spectrum (b)

IV. Analysis of a Stochastic Part of Signal

Extracting mean function from the signal we obtained its stochastic part $\overset{\circ}{\xi}(nh)=\xi(nh)-\hat{m}(nh)$. Its covariance function and spectral density are depicted in Fig. 6.

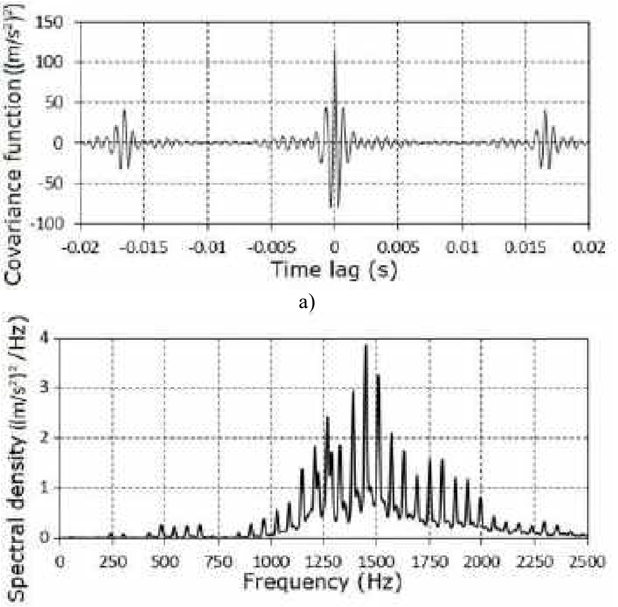
The second order hidden periodicities were detected using the variance functional [15, 27–30]

$$\hat{F}_2(0,\theta)=\frac{1}{2K+1}\sum_{n=-K}^{K}\hat{R}_{\xi}^2(nh,0,\theta) \qquad (6)$$

$$\hat{R}_{\xi}^2(nh,0,\theta)=\sum_{k=1}^{L_2}\begin{bmatrix}\hat{C}_k^{(\xi)}(jh,\theta)\cos k\dfrac{2\pi}{\theta}nh+\\ +\hat{S}_k^{(\xi)}(jh,\theta)\sin k\dfrac{2\pi}{\theta}nh\end{bmatrix},$$

$$\left\{\begin{matrix}\hat{C}_k^{(\xi)}\left(jh,\theta\right)\\ \hat{S}_k^{(\xi)}\left(jh,\theta\right)\end{matrix}\right\}=\frac{2}{2K+1}\sum_{n=-K}^{K}\left[\xi\left(nh\right)-\hat{m}\left(nh\right)\right]\times$$

$$\times\left[\xi\left(\left(n+j\right)h\right)-\hat{m}\left(\left(n+j\right)h\right)\right]\left\{\begin{matrix}\cos k\dfrac{2\pi}{\theta}nh\\ \sin k\dfrac{2\pi}{\theta}nh\end{matrix}\right\}.$$



a)

b)

Fig. 6. Covariance function (a) and power spectral density (b) of the stochastic part of vibration signal.

The estimators of the variance period $\hat{P}$ are found as the maximum point of statistics in (6) with respect to the test period θ. A graph of the dependence of the variance functional (6) on the test frequency is presented in Fig. 7. The maximum point (i.e. estimator for the basic frequency) is equal to 60.420 *Hz*. The variance amplitude spectrum was calculated as

$$\hat{V}\left(k\hat{f}_0\right)=\sqrt{\left[C_k^{(\xi)}\left(0,\hat{P}\right)\right]^2+\left[S_k^{(\xi)}\left(0,\hat{P}\right)\right]^2} \qquad (7)$$

It is presented in Fig. 9. This spectrum is narrower than that of the deterministic oscillations. We may take into consideration only 25–27 harmonics for the variance representation in (7). Since the amplitudes $\hat{V}\left(k\hat{f}_0\right)$ are determined by the joint correlations of the spectrum harmonics with frequencies, shifted by $k\hat{f}_0$ [15, 27–30], the highest frequency for the variance harmonics can't exceed the signal spectrum width (Fig. 6b).

To calculate the variance estimator, following statistics should be used:

$$\hat{b}_{\xi}\left(t,0,\hat{P}\right)=\hat{B}_0^{(\xi)}\left(0\right)+\sum_{k=1}^{L_2}\left[\begin{matrix}\hat{C}_k^{(\xi)}\left(0,\hat{P}\right)\cos k\dfrac{2\pi}{\hat{P}}t+\\ +\hat{S}_k^{(\xi)}\left(0,\hat{P}\right)\sin k\dfrac{2\pi}{\hat{P}}t\end{matrix}\right] \qquad (8)$$

where

$$\hat{B}_0^{(\xi)}\left(0\right)=\frac{1}{2K+1}\sum_{n=-K}^{K}\left[\xi\left(nh\right)-\hat{m}\left(nh\right)\right]^2.$$

If condition $h\leq P/\left(4L_2+1\right)$ is satisfied, formula (8) can be used as interpolation that allows to calculate the variance values for all $t\in\left[0,P\right]$ [15, 27–30].

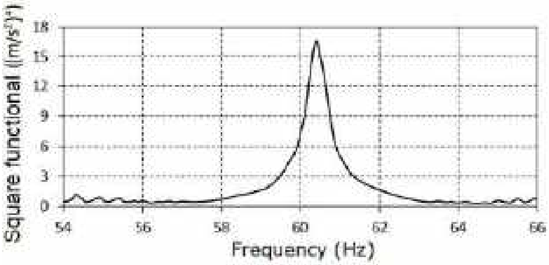


Fig. 7. Dependence of the functional in (6) on the test frequency

The time changes in the variance estimator in (8) are presented in Fig. 8. They have the form of short, powerful pulses that follow one another over the rotation period.

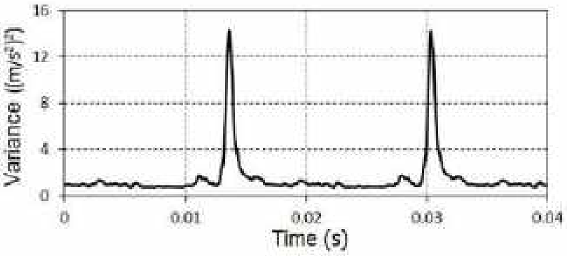


Fig. 8. Variance estimator (9) within one period of hidden periodicity.

The variance time dependence of the (Fig. 8) and its spectrum (Fig. 9) indicate that the mechanism is significantly damaged.

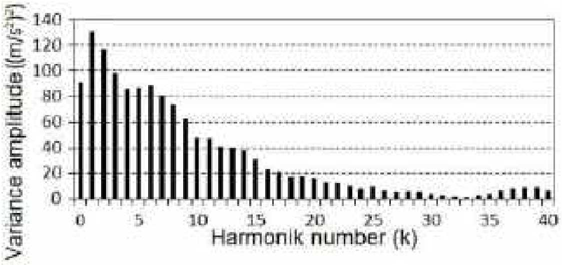


Fig. 9. Amplitude spectrum (11) of variance estimator.

The stochastic signal can be represented by the superposition of the high-frequency narrow-band modulated carrier harmonics. Using of the bandpass filtering and Hilbert transform [27–29] these modulations were analyzed. We first separate three central components around the peak $\lambda_0=1453\ Hz$ (Fig. 10).

Each component is represented by the Rise equations:

$$\xi_0\left(nh\right)=\mu_0^c\left(nh\right)\cos 2\pi\lambda_0 nh+\mu_0^s\left(nh\right)\sin 2\pi\lambda_0 nh\,, \quad (9)$$

$$\xi_1^+\left(nh\right)=\mu_1^c\left(nh\right)\cos 2\pi\lambda_1^+ nh+\mu_1^s\left(nh\right)\sin 2\pi\lambda_1^+ nh\,, (10)$$

$$\xi_1^-\left(nh\right)=\mu_{-1}^c\left(nh\right)\cos 2\pi\lambda_1^- nh+\mu_{-1}^s\left(nh\right)\sin 2\pi\lambda_1^- nh\,,(11)$$

here $\lambda_1^+=\lambda_0+f_0$, $\lambda_1^-=\lambda_0-f_0$.

The covariance functions estimators for components have the form of slowly decaying oscillations (Fig. 11a), and estimators of the spectral density have sharp peaks at frequencies λ_0, λ_1^+ and λ_1^-. The values of the non-zero covariance component estimator for each component are negligible (Fig. 11b, 11c), and can therefore be considered as stationary random processes. However, the results of processing the sum of the processes in (9)–(11) show that these components are jointly periodically nonstationary processes.

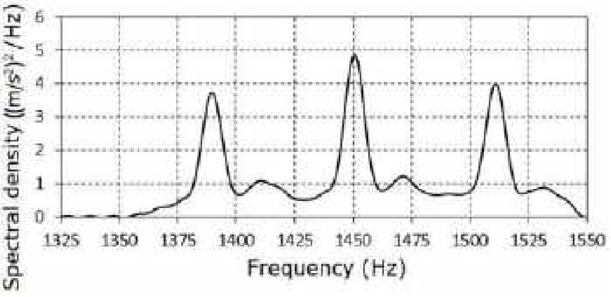


Fig. 10. Power spectral density of the filtered signal

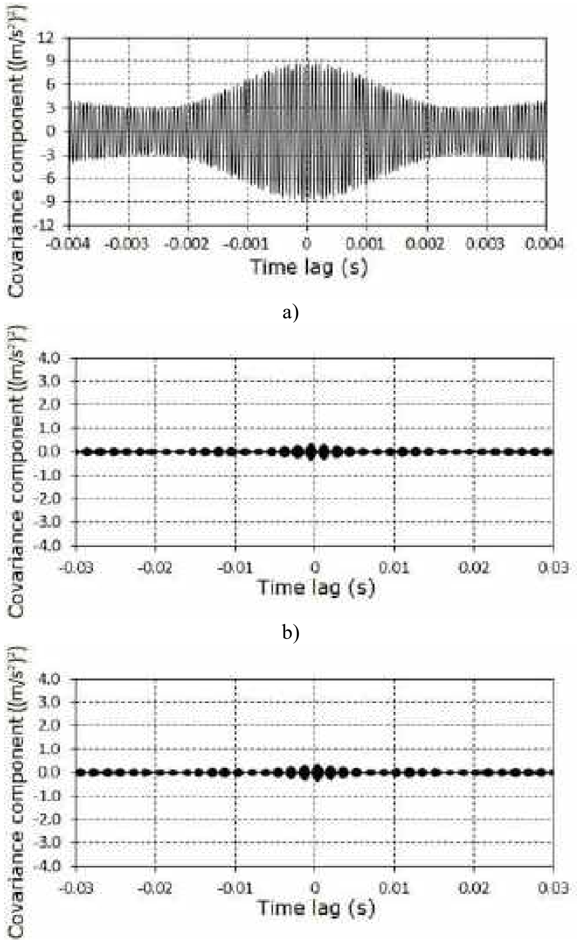


Fig. 11. The covariance components estimators for the one component filtered process $\xi_1^+(nh)$: (a) $\hat{B}_0^{(\xi_1^+)}(u)$, (b) $\hat{C}_1^{(\xi_1^+)}(u)$, (c) $\hat{S}_1^{(\xi_1^+)}(u)$.

The zero^th covariance component of the sum is determined by adding the covariance functions of each component, and it has a group structure (Fig. 12a) that can be explained by the close item frequencies.

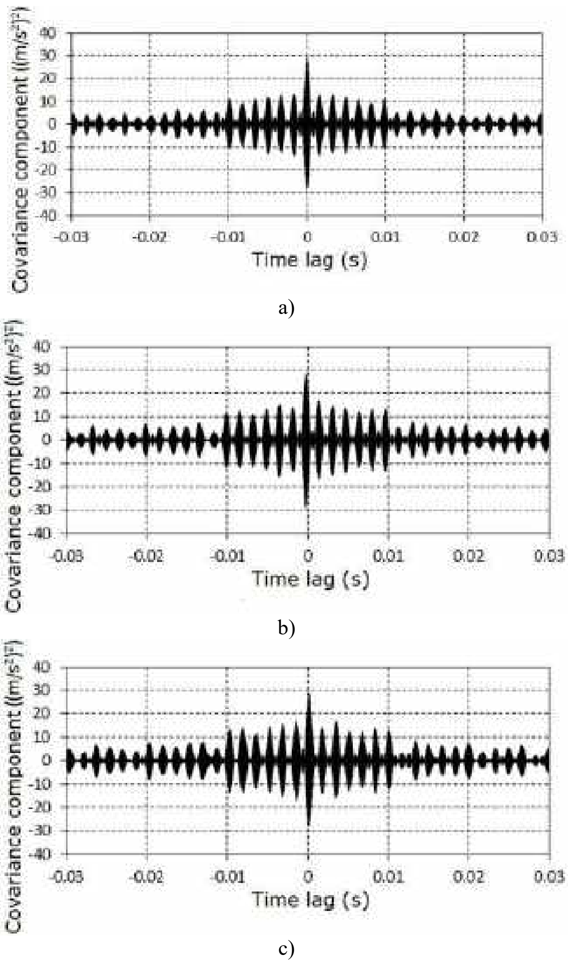


Fig. 12. Estimators of the (a) $\hat{B}_0^{(\xi)}(u)$, (b) $\hat{C}_1^{(\xi)}(u)$, (c) $\hat{S}_1^{(\xi)}(u)$ first covariance components for three-component signal.

The estimators of the first (Fig. 12b, 12c) and second (Fig. 13a, 13b) covariance components have similar forms.

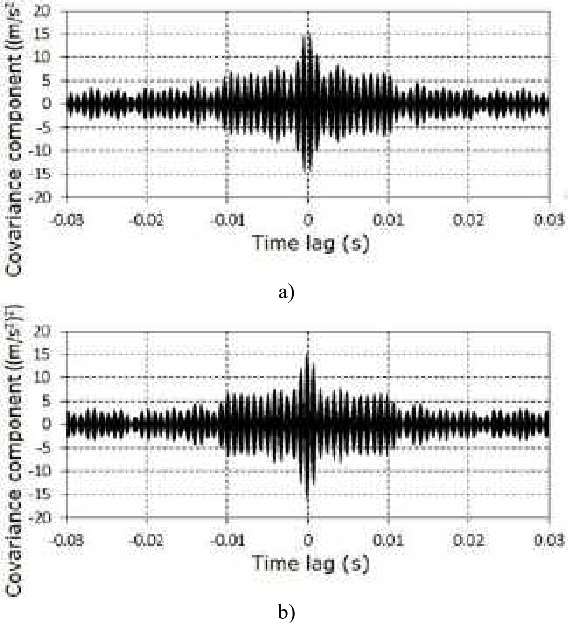
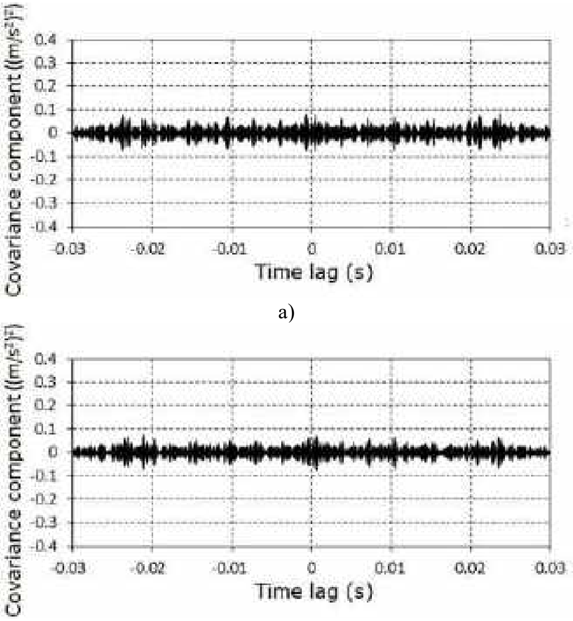


Fig. 13. Estimators of the second cosine $\hat{C}_2^{(\xi)}(u)$ (a) and sine $\hat{S}_2^{(\xi)}(u)$ (b) covariance components.

The first component is determined by the correlations between the processes ξ_0 and ξ_1^+ as well as ξ_0 and ξ_1^-. The correlations of ξ_1^- and ξ_1^+ determines the second component. The values of the third and higher components are negligible (Fig. 14), since the spectrum components of processes, which frequencies of are shifted by $kf_0 > 3$, are practically absent.



a)

b)

Fig. 14. Estimators of the (a) third cosine $\hat{C}_3^{(\xi)}(u)$ (a) and sine $\hat{S}_3^{(\xi)}(u)$ (b) covariance components.

It can be shown that other high-frequency components have close properties.

Conclusion

The covariance structure analysis of real vibration signal based on its model as periodical non-stationary process involving its Hilbert transform was performed. It was shown, that time changes of the signal moment function for the second order are results of the cross-covariances of its narrow-band high-frequency items. To obtain the sensitive indicator for the mechanism condition monitoring we must take into consideration all components which are accentually correlated.

References

[1] P.D. McFadden, J.D. Smith, “Vibration monitoring of rolling element bearings by the high frequency resonance technique – A review”, Tribol. Int., 17, 1984, pp. 3–10.

[2] W.A. Gardner, Cyclostationarity in communications and signal processing, New York: IEEE Press; 1994.

[3] D. Ho, R.B. Randall, Optimization of bearing diagnostic techniques using simulated and actual bearing fault signals”, Mech. Syst. Signal Process. 14 (5), 2000, pp. 763–788.

[4] H. Wang, “Early detection of gear tooth cracking using the resonance demodulation technique”, Mech. Syst. Signal Process., 15(5), 2001, pp. 887–903.

[5] H. Hurd, A. Miamee, Periodically correlated random sequences: Spectral theory and practice, New York: Wiley; 2007.

[6] J. Antoni, R.B. Randall, “A stochastic model for simulation and diagnostics of rolling element bearings with localized faults”, ASME J. Vib. Acoust., 125, 2003, pp. 282–289.

[7] R.B. Randall, J. Antoni, “Rolling element bearing diagnostics – A tutorial”, Mech. Syst. Signal Process., 25, 2011, pp. 485–520.

[8] D. Abboud, M. El Badaoui, W.A. Smith, R.B. Randall, “Advanced bearing diagnostics: A comparative study of two powerful approaches”, Mech. Syst. Sig. Process., 114, 2019, pp. 604–627.

[9] I.Ya. Dolinska, “Evaluation of the Residual Service Life of a Disk of the Rotor of Steam Turbine with Regard for the Number of Shutdowns of the Equipment”, Mater. Sci., 53(5), 2018, pp. 637–644.

[10] C. Peeters, J. Antoni, J. Helsen, “Blind filters based on envelope spectrum sparsity indicators for bearing and gear vibration-based condition monitoring”, Mech. Syst. Signal Process., 138, 2020, pp. 106556.

[11] H. Konstantin-Hansen, “Envelope analysis for diagnostics of local faults in rolling element bearings”, Denmark: Bruel & Kjaer Application Note, BD0501, 2003.

[12] Y. Xu, D. Zhen, X. Gu et al., “Autocorrelated envelopes for early fault detection of rolling bearings”, Mech. Syst. Sig. Process., 146, 2021, 106990.

[13] J. Antoni, “Cyclostationarity by examples”, Mech. Syst. Signal Process., 23, 2009, pp. 987–1036.

[14] I. Javorskyj, R. Yuzefovych, O. Lychak, R. Slyepko, P. Semenov, “Detection of distributed and localized faults in rotating machines using periodically non-stationary covariance analysis of vibrations”, Meas. Sci. Technol., 34, 2023, 065102.

[15] I. Javorskyj, I. Matsko, R. Yuzefovych, O. Lychak, R. Lys, “Methods of hidden periodicity discovering for gearbox fault detection”, Sensors, 21, 2021, 6138.

[16] A. Napolitano, Cyclostationary processes and time series: Theory, applications, and generalizations, Elsevier, Academic Press, 2020.

[17] S. Tyagi, S.K. Panigrahi, “An improved envelope detection method using particle swarm optimisation for rolling element bearing fault diagnosis”, J. Comput. Des. Eng., 4, 2017, pp. 305–317.

[18] A. Mauricio, W.A. Smith, R.B. Randall, J. Antoni, “Improved Envelope Spectrum via Feature Optimisation-gram (IESFOgram): A novel tool for rolling element bearing diagnostics under non-stationary operating conditions”, Mech. Syst. Sig. Process., 144, 2020, 106891.

[19] N. Sawalhi, R.B. Randall, H. Endo, “The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis”, Mech. Syst. Signal Process., 21, 2007, pp. 2616–2633.

[20] A.C. McCormick, A.K. Nandi, “Cyclostationarity in rotating machine vibrations”, Mech. Syst Signal Process 12(2), 1998, pp. 225–242.

[21] C. Capdessus, M. Sidahmed, J.L. Lacoume, “Cyclostationary processes: Application in gear fault early diagnostics”, Mech. Syst. Signal Process., 14, 2000, pp. 371–385.

[22] I. Antoniadis, G. Glossiotis, “Cyclostationary analysis of rolling-element bearing vibration signals”, J. Sound Vib., 248, 2001, pp. 829–845.

[23] J. Antoni, P. Borghesani, “A statistical methodology for the design of condition indicators”, Mech. Syst. Signal Process., 114, 2019, pp. 290–327.

[24] E. Bedrosian, “A product theorem for Hilbert transforms”, Proceedings of the 51th IEEE, 1963, pp. 868–869.

[25] D. Wang, X. Zhao, L.-L. Kou, Y. Qin, Y. Zhao, K.-L. Tsui, “A simple and fast guideline for generating enhanced/squared envelope spectra from spectral coherence for bearing fault diagnosis”, Mech. Syst. Signal Process., 122, 2019, pp. 754–768.

[26] J. Antoni, F. Bonnardot, A. Raad, M. El Badaoui, “Cyclostationary modeling of rotating machine vibration signals”, Mech. Syst. Signal Process., 18, 2004, pp. 1285–1314.

[27] I. Javorskyj, R. Yuzefovych, O. Lychak, P. Kurapov, “Hilbert transform for analysis of daily changes of the Earth magnetic field”, IEEE 12th International Conference on Electronics and Information Technologies, 2021, pp. 181–185.

[28] I. Javorskyj, R. Yuzefovych, O. Lychak, R. Sliepko, P. Semenov, “Hilbert transform for analysis of amplitude modulated wide-band random signals”, XIIth International Conference on Advanced Computer Information Technologies, 2022, pp. 68–71.

[29] I. Javorskyj, R. Yuzefovych, O. Lychak, R. Sliepko, M. Varyvoda, “Hilbert transform of periodically non-stationary random signals: narrow-band high frequency amplitude modulation”, 2022 IEEE 3rd KhPI Week on Advanced Technology, KhPI Week 2022, 183771.

[30] I. Javorskyj, R. Yuzefovych, O. Lychak, I. Matsko, P. Semenov, “Evaluation of the mechanism damage using model of vibration signal as a periodically correlated random process”, Procedia Structural Integrity, 36, 2022, pp. 122–129.

The Accuracy of Speech Transmission Index Estimation under Conditions of Joint Action of Noise and Reverberation

Arkadiy Prodeus
Acoustic and Multimedia Electronic Systems Department
NTUU “Igor Sikorsky Kyiv Polytechnic Institute”
Kyiv, Ukraine
aprodeus@gmail.com

Oleksandr Dvornyk
Acoustic and Multimedia Electronic Systems Department
NTUU “Igor Sikorsky Kyiv Polytechnic Institute”
Kyiv, Ukraine
alexanderdvornyk@gmail.com

Anton Naida
Acoustic and Multimedia Electronic Systems Department
NTUU “Igor Sikorsky Kyiv Polytechnic Institute”
Kyiv, Ukraine
naida.a.s.2001@gmail.com

Maryna Didkovska
Department of Mathematical Methods of System Analysis
NTUU “Igor Sikorsky Kyiv Polytechnic Institute”
Kyiv, Ukraine
maryna.didkovska@gmail.com

***Abstract* — In this paper, the full modulation method of speech intelligibility index (STI) estimation and its modification in the form of the full formant-modulation (FM) method are compared in terms of measurement accuracy in conditions where the speech signal is masked by noise and reverberation. Dependences of STI estimation errors on signal-to-noise ratios and on the duration of test signals for the reverberation time typical for university auditoriums of 0.8 s were obtained. It is shown that the accuracy of STI estimation in the presence of reverberation practically does not depend on the choice of estimation method. The obtained results indicate that an acceptable for practical use error of 0.01-0.02 of STI estimation in the conditions of joint action of noise and reverberation can be ensured when using test signals lasting 8-16 s.**

Keywords* — speech transmission index, full modulation method, full formant-modulation method, fast method, estimation error

I. Introduction

Since noise and reverberation interferences are always present in speech information transmission channels, the task of evaluating speech intelligibility in such channels is relevant [1], [2], [3], [4]. It is known that the formant method [1], [5] is the best in the conditions of the exclusive effect of noise interference, as it allows to achieve the maximum accuracy of the articulation index (AI) with the minimum measurement time [6]. A significant advantage of the modulation method [7] is the possibility to take into account the distortion of the speech signal not only by noise, but also by reverberation. The FM method [6] is a type of modulation method, so it also allows taking into account the combined effect of noise and reverberation on speech intelligibility. An additional possibility of calculating the AI articulation index can be attributed to the advantage of the FM method.

There are two versions, full and fast, of modulation and FM methods [6], [7], [8]. The disadvantage of the full version is the significant duration (up to 16 minutes) of the measurement procedure, caused by the need to use a large number (almost 100) of rather long (about 10 s) test signals. Therefore, simplified fast versions such as RASTI, STIPA and STITEL [7], [8] are often used in practice, where only one test signal lasting 15-20 s is used.

At the same time, in some modern hardware and software measurement systems [9], the possibility of performing STI measurement using the full modulation method has been realized. However, it is rather difficult to find a justification for choosing the above values of the duration T of the test signals for the full modulation method in the literature. For the formant-modulation method, such justification is also unknown. The purpose of this article is to eliminate this shortcoming.

II. Problem Statement

In the full modulation method, 14 test signals are used to measure STI [6]

$$x_i(t) = \xi(t)\sqrt{f_i(t)}\ ,\ \ f_i(t) = 1 + \sin 2\pi F_i t\ ,\ \ i = \overline{1,14}\ , \quad (1)$$

$$F_i = 0.63,\ 0.8,\ 1,\ 1.25,\ 1.6,\ 2,\ 2.5,\ 3.15,\ldots$$

$$\ldots 4,\ 5,\ 6.3,\ 8,\ 10,\ 12.5\ \text{Hz},$$

$\xi(t)$ is a stationary noise with speech spectrum, $f_i(t)$ is a modulation function, F_i is a modulation frequency. Signals $x_i(t)$ are emitted in turn by a sound source located at the point where the speaker is usually located. At the point where the listener is located, a signal $y_i(t) = x_i(t) \otimes h(t) + n(t)$ is received by the microphone, $n(t)$ is the noise with a certain signal-to-noise ratio (SNR), $h(t)$ is the room impulse response (RIR), $\otimes$ is the convolution symbol. The signals $y_i(t)$ are then filtered by a 7-band octave filter bank, resulting in a set of 98 signals $y_{k,i}(t),\ k = \overline{1,7},\ i = \overline{1,14}$.

The STI calculation algorithm is given in [7]. The key point in this algorithm is the calculation and use of an effective signal-to-noise ratio

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

$$SNR_{eff\,k,i} = 10\lg\frac{m_{k,i}}{1-m_{k,i}}, \qquad (2)$$

$$\tilde{m}_{k,i} = \frac{2\,|A_{k,i}(F_i)|}{|A_{k,i}(0)|}, \quad A_{k,i}(F_i) = \frac{1}{T}\int_0^T y_{k,i}^2(t)e^{-j2\pi F_i t}dt$$

$|\cdot|$ is module symbol, ~ is estimate symbol.

In the next step, the $SNR_{eff\,k,i}$ values are subjected to a non-linear transformation to obtain the modulation transfer index

$$MTI_k = \frac{1}{14}\sum_{i=1}^{14} TI_{k,i}, \qquad (3)$$

$$TI_{k,i} = \begin{cases} \dfrac{SNR_{eff\,k,i}+15}{30}, & -15 < SNR_{eff\,k,i} < 15; \\ 0, \quad SNR_{eff\,k,i} \le -15; & \\ 1, \quad SNR_{eff\,k,i} \ge 15; & \end{cases},$$

The FM method algorithm proposed in [6] differs from the algorithm of the modulation method only in calculating, after $SNR_{eff\,k,i}$ calculation according to (2), the modulation transfer index in another way:

$$MTI_k = \begin{cases} \dfrac{E_k+15}{30}, & -15 < E_k < 15; \\ 0, \quad E_k \le -15; & \\ 1, \quad E_k \ge 15; & \end{cases}, \qquad (4)$$

$$E_k = \frac{1}{14}\sum_{i=1}^{14} SNR_{eff\,k,i}.$$

Given the similarity of the algorithms of the modulation and FM methods, it can be expected that the corresponding STI estimates will be close. Since the validity of this assumption has not yet been confirmed, the purpose of this paper is to fill this gap.

III. Set up of the Study

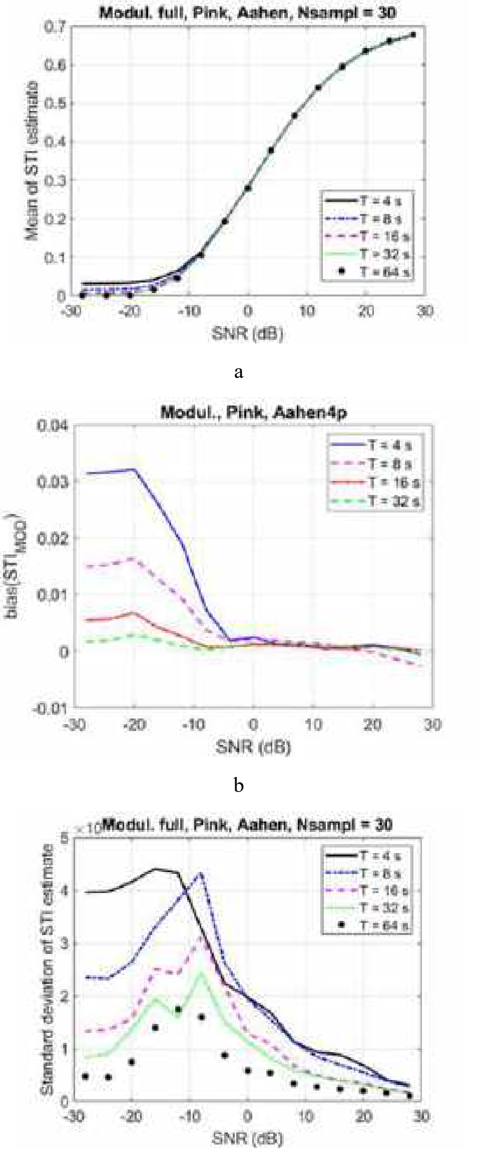
Signals $y_i(t) = x_i(t) \otimes h(t) + n(t)$ were generated with a sampling frequency of 22050 Hz by computer simulation. The record of the RIR $h(t)$ of the real university auditorium with a volume of 370 m^3 and a reverberation time of 0.8 s was used [10]. For the study, pink noise $n(t)$ was used as it corresponds better to real situations than white noise. Calculations of STI by each of the methods were performed separately. For each combination of *SNR* and *T* parameters, 30 STI estimates were calculated, which made it possible to estimate the mathematical expectation, bias, and standard deviation of the STI estimates with sufficient accuracy for engineering applications.

The software proposed in [6] was modified for STI calculations using the full modulation method and used for STI calculations using the full FM method. Calculations were performed in the Matlab R2022a environment.

IV. Results of the Study

A. Full Modulation Method

The results of estimating, based on 30 samples, the mathematical expectation, bias (relative to the case *T*=64 s) and standard deviation of STI estimates for the full modulation method are shown in Fig. 1.



a

b

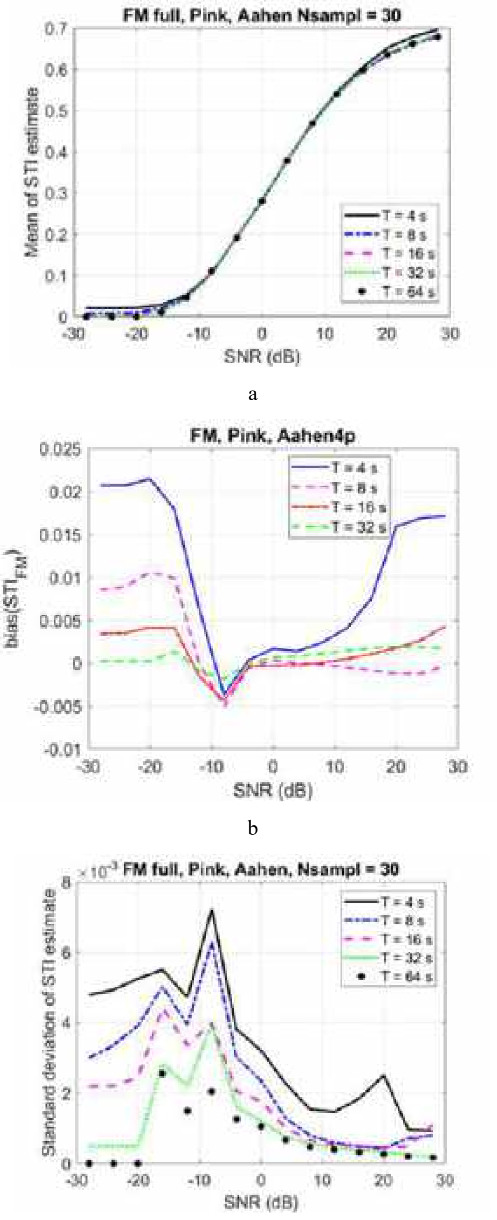
c

Fig. 1. Full modulation method: estimates of expectation (a), bias (b) and standard deviation (c)

Estimates of expectation (Fig. 1a) indicate the presence of a bias in STI estimates (Fig. 1b), which decreases with *T* and *SNR* increasing. To obtain quantitative value of the bias, the STI estimate for the case *T*=64 s was used as a true value.

B. Full FM Method

The results of estimating the expectation, bias (relative to the case *T*=64 s) and standard deviation of STI estimates for the full FM method are shown in Fig. 2.



a

b

c

Fig. 2. Full FM method: estimates of expectation (a), bias (b) and standard deviation (c)

C. Comparison of Modulation and FM methods

The above results indicate a significant similarity, in terms of the accuracy of STI measurements, of the full modulation and full FM methods. In order to compare the STI estimates for these methods more clearly, the differences in the expectation estimates

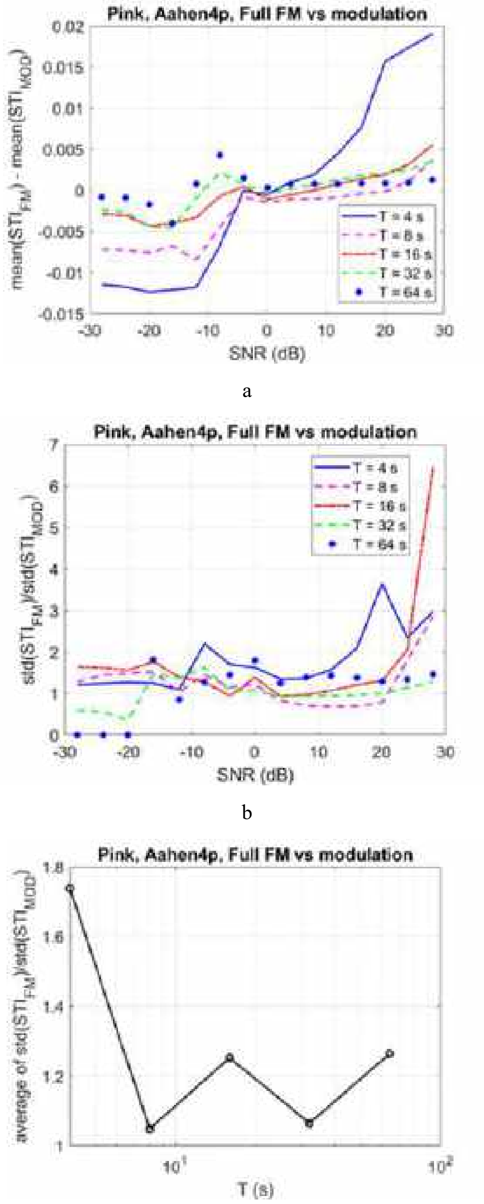
$$\Delta_{FM,mdl} = \overline{STI}_{FM} - \overline{STI}_{mdl} \tag{5}$$

and the ratio of the standard deviation estimates

$$\Lambda_{FM,mdl} = \overline{\sigma STI}_{FM} \Big/ \overline{\sigma STI}_{mdl} \tag{6}$$

were calculated, where $\overline{STI}_{FM}$ and $\overline{STI}_{mdl}$ are the average values of STI estimates obtained by full FM and modulation methods, respectively, $\overline{\sigma STI}_{FM}$ and $\overline{\sigma STI}_{mdl}$ are estimates of the corresponding standard deviations.

The results of calculations according to (5) and (6) are shown in Fig. 3a and 3b, respectively. The results of averaging the values (6) in the interval *SNR*=-20...20 dB for different *T* are shown in fig. 3c.



a

b

c

Fig. 3. Comparison of the methods: difference in expectation (a), ratio of standard deviations (b), average of standard deviations ratio (c)

V. Discussion

As can be seen in Fig. 1b, the STI estimate obtained by the full modulation method is biased towards higher values at *SNR* < -5 dB. At *SNR* > -5 dB, the STI estimate is significantly less biased, and the amount of bias decreases with *SNR* increasing. For *T* = 16 s, the value of the shift does not exceed 0.007 in the range of *SNR* = -28…+28. The standard deviation of the STI estimate (Fig. 1c) has a maximum in the range of *SNR* values from -18 dB to -8 dB and decreases to very small values as the *SNR* approaches 28 dB. A noticeable decrease also occurs as the *SNR* approaches minus 28 dB. For *T* = 16 s, the value of the standard deviation does not exceed 0.003 in the range of *SNR* = -28…+28.

As can be seen in Fig. 2a and 2b, the STI estimate obtained by the full FM method is noticeably biased towards higher values at *SNR* < -10 dB and *SNR* > +10 dB. In general, the bias value decreases with increasing duration. For *T* = 16 s, the value of the bias does not exceed 0.004 in the range of *SNR* = -28…+28. The standard deviation of the STI estimate (Fig. 2c) has a maximum in the range of *SNR* values from minus 18 dB to minus 8 dB and decreases to very small values as the *SNR* approaches 28 dB. Some reduction in the standard deviation of the STI score also occurs as the *SNR* approaches minus 28 dB. For *T* = 16 s, the value of the standard deviation does not exceed 0.004 in a wide range of *SNR* = -28…+28.

Shown in Fig. 3a results show that at *T* = 16 s, the modulus of the difference between the average values of STI estimates does not exceed 0.005 in a wide range of *SNR* = -28…+28 dB. As can be seen in Fig. 3b, the results of calculations according to (6) show that the standard deviations of the estimates also differ little. The deviations ratio is close to 1 in the range of *SNR* = -28…+15 dB. However, for *SNR* > 15 dB and *T* < 32 s, the standard deviation of the FM method is 3-6 times higher than that of the modulation method. A graph of the dependence of the averaged values of the ratio (6) on the duration *T* of the test signals (1) over the range of *SNR* = -20…+20 dB is shown in Fig. 3c. It can be seen that the standard deviations of STI estimates for the FM method are only about 20% higher than those for the modulation method, although in the case of *T* = 4 s such an excess is more significant and reaches 80%.

The values of the maximum STI estimation errors obtained in this paper are presented in the Table 1, where Δ is the maximum bias, within the interval -28 dB < *SNR* < 28 dB, Σ is the maximum standard deviation, $\Omega = \sqrt{\Delta^2 + \Sigma^2}$ is the maximum total measurement error.

TABLE I. MEASUREMENT ERRORS

Method	***T*** **(s)**	Δ	Σ	Ω
Modulation	**4**	0.032	0.004	0.032
	8	0.016	0.004	0.016
	16	0.007	0.003	0.008
	32	0.003	0.002	0.004
	64	0	0.001	0.001
FM	**4**	0.022	0.007	0.023
	8	0.011	0.006	0.013
	16	0.004	0.004	0.006
	32	0.002	0.004	0.004
	64	0	0.002	0.002

Since the value of just noticeable difference JND = 0.03 [11] is considered acceptable for practical application as the STI estimation error, it can be seen from the Table 1 that this requirement is practically satisfied even at *T* = 4 s. The maximum total error of STI estimation is close to 0.03 for both methods in this case. At *T* = 8 s and *T* = 16 s, the maximum total STI estimation error for both methods is close to 0.02 and 0.01, respectively. Note that these results are in good agreement with [6], where it is indicated that the STI estimation error is 0.02 for a test signal duration of 10 s.

However, it should be noted that this conclusion is actually verified for the situation when the reverberation time in the room does not exceed *T60* = 1 s. For *T60* > 1 s, this conclusion needs additional verification.

CONCLUSION

Under the combined effect of noise and reverberation, the full modulation and full FM STI measurement methods provide virtually the same accuracy in the range of signal-to-noise ratios from minus 28 dB to plus 28 dB and in the range of test signal durations from 4 s to 64 s, provided that the time reverberation does not exceed 1 s. At the same time, the duration of *T* = 4 s of test signals is minimally acceptable and provides an estimation error close to 0.03. The use of test signals with a duration of *T* = 8 s and *T* = 16 s allows to reduce the STI estimation error to 0.02 and 0.01, respectively.

The case when the reverberation time exceeds 1 s requires additional research.

REFERENCES

[1] J. Collard, “Theoretical Study of the Articulation and Intelligibility of a Telephone Circuit,” Electrical Communication, vol.7, 1929, p. 168.

[2] K. Kryter, The Effects of Noise on Man, Academic Press, New York and London, 1970, 612 p.

[3] H. Steeneken, T. Houtgast, "A physical method for measuring speech-transmission quality,” J.Acoust. Soc. Am., 67, 1980, pp. 318-326.

[4] K. Rhebergena, “A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners,” J. Acoust. Soc. Am., 117 (4), Pt. 1, April 2005, pp. 2181-2192.

[5] ANSI S3.5-1969, American National Standard, "Methods for Calculation of the Articulation Index," American National Standards Institute New York, 1969.

[6] A. Prodeus, “Formant-Modulation Method of Speech Intelligibility Evaluation: Measuring and Exactness,” Proc. VII Int. Conf. MEMSTECH 2011, Lviv, Polyana, Ukraine, 2011, pp.54-60.

[7] British Standard BS EN 60268-16. Sound system equipment. Part 16. Objective rating of speech intelligibility by speech transmission index. 2011.

[8] A. Prodeus, "Rapid version of a formant-modulation method of speech intelligibility estimation," Proc. VII Int. Conf. MEMSTECH 2011, Lviv, Polyana, Ukraine, 2011, pp. 61-63.

[9] NTi Audio, Application note. Speech Intelligibility. Measurement with the XL2 analyzer. Dec. 2020. 28 p.

[10] M. Jeub, M. Schafer, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” Proc. Int. Conference on Digital Signal Processing (DSP), Santorini, Greece, 2009.

[11] J. Bradley, R. Reich, R. Norcross, “A just noticeable difference in C50 for speech”, Applied Acoustics, (58), 1999, pp. 99-108.

The Impact of Parameters on the Efficiency of Keypoints Detection and Description

Andriy Fesiuk
department of electronics and computer technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
andrii.fesiuk@lnu.edu.ua

Yuriy Furgala
department of electronics and computer technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
yuriy.furhala@lnu.edu.ua

***Abstract* — This paper comprehensively investigates the efficiency and performance of the keypoints detection and description methods in computer vision and image processing. Four widely used methods - SIFT, SURF, ORB, and BRISK are compared and analyzed. The study aims to assess the influence of different parameters on detecting and describing keypoints, specifically focusing on the limitation of the number of keypoints and Lowe's ratio values. The effectiveness of each method is evaluated by analyzing the similarity coefficient for pairs of images with varying degrees of similarity. It has been found that the similarity coefficient is also rising, with Lowe's ratio rising. At the same time, both Lowe's ratio and detected keypoints number limitation significantly impact the achieved methods' performance. The results obtained from the experiments provide valuable insights into the pros and cons of each keypoints method. Furthermore, we suggest how to optimize parameter settings to achieve optimal performance. The findings of this research can benefit researchers and developers working in computer vision and image processing.**

Keywords* — *keypoints, detection, description, matching, SIFT, SURF, ORB, BRISK.

I. Introduction

Computer vision and digital image processing have witnessed significant advancements in recent years, driven by the increasing availability of visual data and the demand for efficient algorithms. In this context, detecting and describing keypoints have become crucial tasks for image analysis and pattern recognition [1-3].

Several keypoints detection and description methods have appeared, each with unique strengths and limitations. Prominent among these methods are SIFT (Scale-Invariant Feature Transform), SURF (Speeded-Up Robust Features), ORB (Oriented FAST and Rotated BRIEF), and BRISK (Binary Robust Independent Elementary Features) [4-9]. These methods have applications in diverse fields, from medical imaging and robotics to surveillance and agriculture [10-15].

Despite their wide adoption, there remains a need to explore the influence of different parameters on the effectiveness of these keypoints methods. The main objective of this study is to investigate the impact of keypoints number limitation and Lowe's ratio test values on accuracy and efficiency.

Additionally, an analysis and comparison of the performance of SIFT, SURF, ORB, and BRISK methods under different settings, focusing on their ability to detect similar images efficiently, was conducted. Our findings will provide valuable insights into the strengths and weaknesses of each method and offer an approach for parameter optimization to achieve satisfying results in practical applications.

II. Methods and Materials

We prepared a dataset of 100 images of beer cans to evaluate the performance of the methods for keypoints detection and description (Fig. 1). The images were taken on the iPhone XS primary camera. The resulting images were in HEIC format in 2268x4032 resolution. Due to the inability of the OpenCV to work with this format directly, the images were converted into PNG format. At the same time, we are focused on the mid-resolution photos at the current stage of our investigation. Hence, the images were downscaled to be in 720x1280 resolution. The dataset falls into ten groups, each containing ten similar images.

Fig. 1 Some samples of investigated images

Each method's implementation for keypoints detection and description was obtained from the OpenCV library [16]. A custom Python program was designed to analyze the effectiveness of each method with different parameter values.

Various parameter settings were explored to investigate the impact of different parameters' values on the detected and described keypoints number limitation [16]. For SIFT and ORB, the "nfeatures" parameter varied from 100 to 1000 with a step size 100. For SURF, the "hessianThreshold" parameter was adjusted from 300 to 600 with step size 50, 600 to 1500 with step size 200, and 1500 to 2700 with step size 300. BRISK's "thresh" parameter varied from 46 to 88 with a step size 3.

Our implementation enabled us to explore various values of Lowe's ratio [4], a key parameter used to filter suitable matches after keypoints detection and description. The program uses Lowe's ratio with varying values from 0.6 to 0.85 with a step size of 0.05. Besides, to compare images based on their keypoints, a matching technique using the "cv2.BFMatcher" method was employed, specifically the "bf.knnMatch" function with parameter k set to 2 [16]. To assess the effectiveness of every method for each pair of images, a similarity coefficient was found, which shows the percentage of matching key points between two images relative to the total number of key points on them.

III. Results and Discussion

Boxplots were utilized to represent valuable information about the spread and variation of the similarity coefficients

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

within each method to investigate the impact of different parameter values thoroughly. These boxplots were categorized into two groups: similar and different. Additionally, the boxplots illustrate the distribution of similarity coefficients for different parameter values of Lowe's ratio and the number of detected keypoints. Also, each graph contains the average value line plots of the similarity coefficient and its standard deviation in the backgrounds.

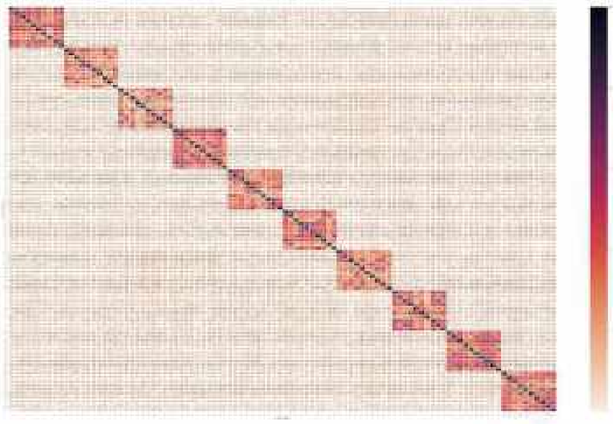


Fig. 2. The similarity matrix for the set of keypoints generated using the SIFT method.

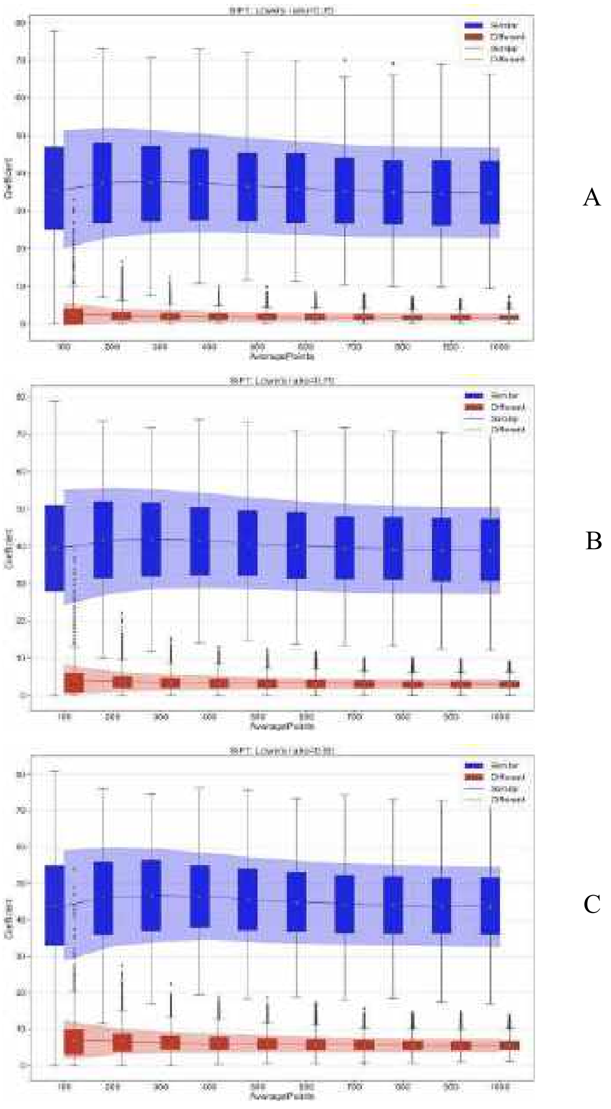


Fig. 3. Results, obtained by SIFT with different Lowe's Ratio (A-0.7, B-0.75, C-0.8) and different number of average points.

A heat map matrix of 100*100 was created for each of the algorithms, corresponding to the number of images. Fig. 2 presents an example of analyzed heatmaps. They contain similarity coefficients ranging from 0 to 100 in the appropriate color for better visual understanding. The diagonal elements represent the similarity of the image to itself and are always equal to 100%.

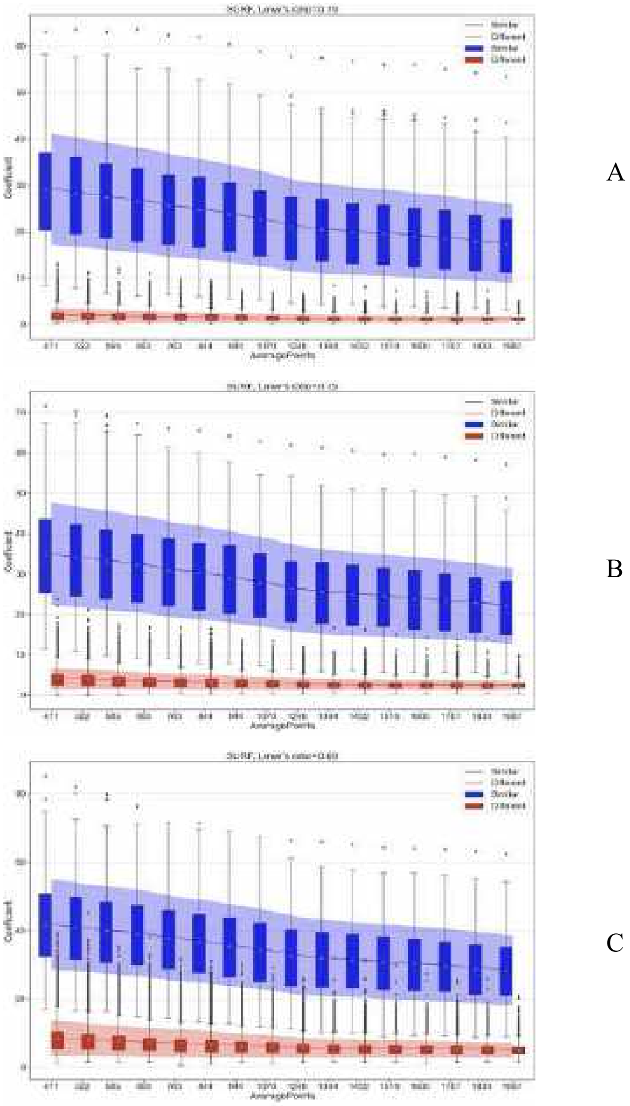


Fig. 4. Results, obtained by SURF with different Lowe's Ratio (A-0.7, B-0.75, C-0.8) and different number of average points.

Fig. 3 depicts the distribution of similarity coefficients for SIFT with varying Lowe's ratio values and different numbers of keypoints. It is observed that increasing Lowe's ratio leads to a decrease in the number of outliers for both similar and different images. Additionally, as the number of detected keypoints increases, the number of outliers decreases, but it slightly rises for keypoints in the range of 700 to 800.

Further analysis reveals that the average and standard deviation ratio decreases with increasing Lowe's ratio. Additionally, as the number of detected keypoints increases, this ratio decreases for similar and different images. The optimal values for SIFT are considered to lie within the ranges of 0.75 to 0.8 for Lowe's ratio and 400 to 1000 for the average number of detected keypoints.

Fig. 4 illustrates the distribution of similarity coefficients obtained by the SURF method. The results show that the similarity coefficients decrease as the number of keypoints increases with different Lowe's ratio values. Furthermore, increasing Lowe's ratio leads to a decrease in the number of outliers for similar images and an increase in outliers for different images. Meanwhile, as the number of detected keypoints increases, the number of outliers decreases for similar and different images.

The ratio between the average and standard deviation decreases overall with increasing Lowe's ratio. Moreover, with an increase in the number of detected keypoints, the ratio increases for similar images but decreases for different images. The considered optimal values for SURF are 0.7 to 0.75 for Lowe's ratio and 470 to 762 for the average number of detected keypoints.

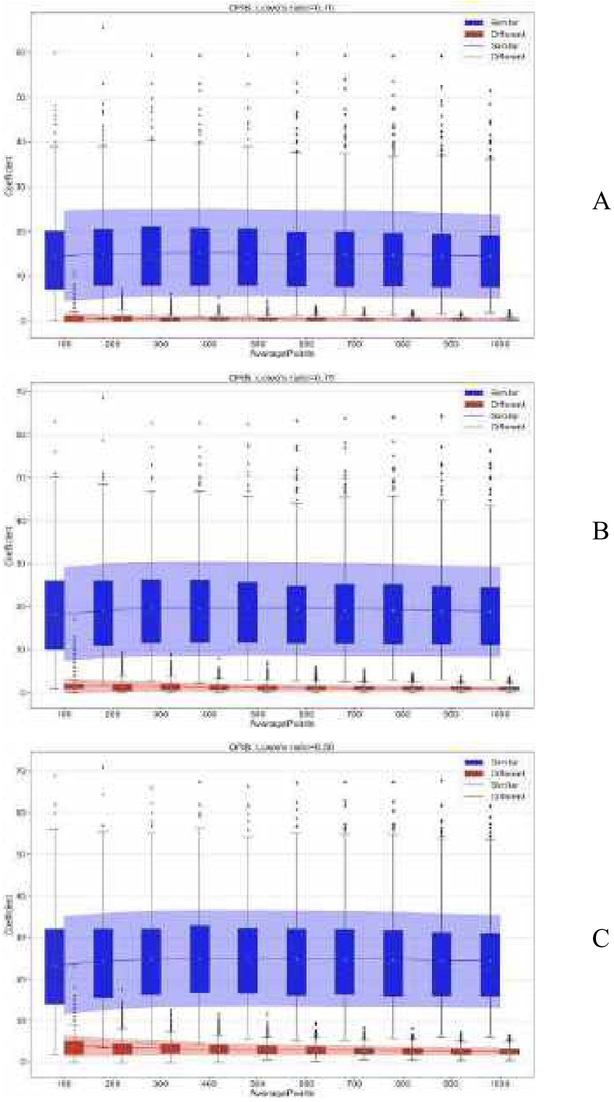


Fig. 5. Results, obtained by ORB with different Lowe's Ratio (A-0.7, B-0.75, C-0.8) and different number of average points.

Fig. 5 demonstrates the results obtained from the ORB method. The similarity coefficients decrease with an increase in the number of detected keypoints. Moreover, the number of outliers decreases with an increase in Lowe's ratio. The behavior of outliers for similar and different images shows fluctuations with the same amplitude, except for a significant fall at the start for different images.

The ratio between the standard and average deviation decreases with an increase in Lowe's ratio. Additionally, as the average number of detected keypoints rises, the ratio decreases for similar and different images. The optimal values for ORB are considered to lie within the ranges of 0.75 to 0.8 for Lowe's ratio and 800 to 1000 for the average number of detected keypoints.

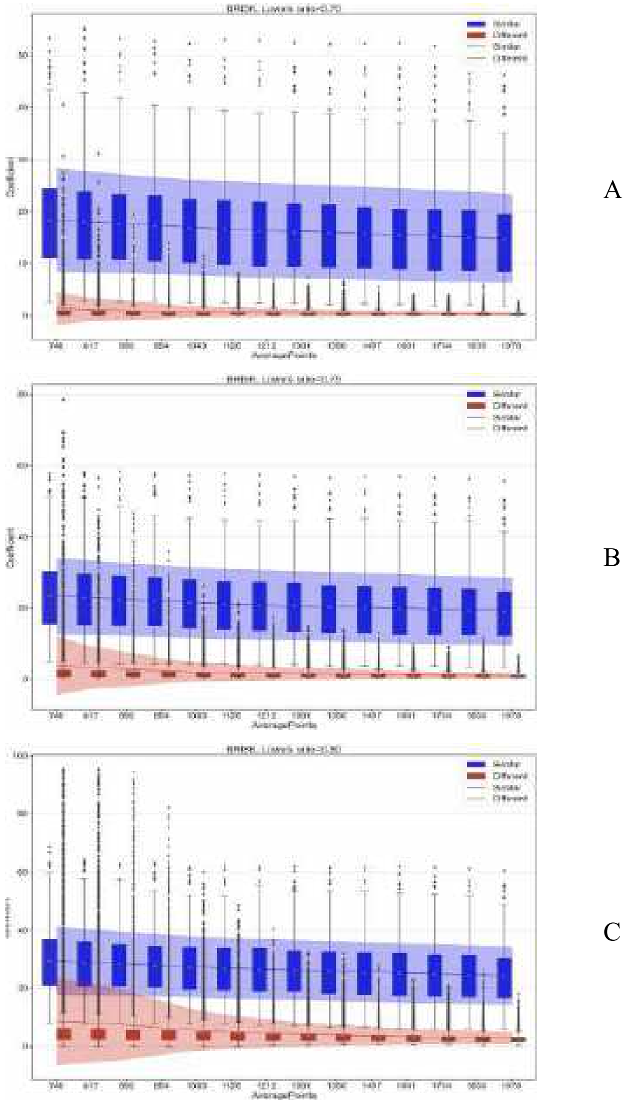


Fig. 6. Results, obtained by BRISK with different Lowe's Ratio (A-0.7, B-0.75, C-0.8) and different number of average points.

Fig. 6 presents the results of the BRISK method. The similarity coefficients increase with the rising number of detected keypoints. With an increase in Lowe's ratio, the number of outliers decreases for similar and different images. However, for similar images, the number of outliers shows a fluctuating pattern with increasing and decreasing amplitude.

The ratio between the standard and average deviation decreases with a decrease in Lowe's ratio. Simultaneously, increasing the number of detected keypoints increases the ratio for similar images but decreases for different images. The optimal values for BRISK are considered to lie within the ranges of 0.7 to 0.75 for Lowe's ratio and 1496 to 1969 for the average number of detected keypoints.

Table 1 presents the time consumed by each method for completion. Notably, SIFT exhibited the longest processing time, surpassing the following method by nearly threefold. The second most time-consuming method was SURF. Conversely, ORB demonstrated the shortest processing time among the evaluated methods. BRISK, the second fastest method, required approximately twice as much time as ORB for completion.

TABLE I. COMPARISON OF DETECTION AND DESCRIPTION SPEED

Methods name	Parameters' value	Average of detecting time, sec
SIFT	nfeatures: 100-1000	12.45
SURF	hessianTheshold: 300-2700	4.75
ORB	nfeatures: 100-1000	1.17
BRISK	tresh: 46-88	1.88

The results showed that SIFT exhibited the highest precision among the methods, making it suitable for tasks prioritizing accurate image recognition. However, it also demonstrated the slowest execution time, which could be a limitation in real-time applications. On the other hand, ORB demonstrated the fastest execution time but showed the lowest precision, making it more suitable for applications that prioritize processing speed over precision.

SURF and BRISK provided a balanced trade-off between precision and speed. While SURF was slightly slower than BRISK, it offered better recognition performance. These methods could be considered viable for applications compromising speed and accuracy.

CONCLUSION

The investigation into the influence of parameter values, particularly Lowe's ratio and keypoints limitation, demonstrated its significant impact on the detection and description methods' performance. Adjusting Lowe's ratio proved essential in filtering detected keypoints and improving matching results. At the same time, keypoints limitation notably impacts the overall similarity coefficient.

All methods demonstrate an increase in similarity coefficient with an increase in Lowe's ratio. As the Lowe's ratio increased, the average and standard deviation ratio generally decreased, indicating a more stable and reliable matching performance. However, the behavior of the ratio concerning the number of detected keypoints varied among the methods. For some methods, the ratio decreased with an increase in the number of keypoints, while for others, it exhibited fluctuations with specific amplitude ranges. Examining optimal parameter values for each method allows us to make informed decisions when choosing the most suitable configurations for specific tasks.

We have acknowledge that the study was limited to a specific dataset of beer can images taken almost in ideal conditions. However, in real-world applications, images may have visual noise or artifacts. Hence, attention should be paid to estimating the performance of the applied methods with distorted images. Furthermore, in future research, alongside with similarity coefficient, we will consider some of the standard metrics, such as SSIM, MS-SSIM, UOI, VIF, and others for method effectiveness evaluation.

In conclusion, this research contributes to the field of computer vision by offering a comparative analysis of keypoints detection and description methods and their performance on a specialized dataset. Future research built upon these findings can explore the methods' performance in different contexts and enhance the efficiency and accuracy of keypoint-based image processing techniques.

REFERENCES

[1] M. Hassaballah, A. A. Abdelmgeid, and H. A. Alshazly, “Image Features Detection, Description and Matching,” in Image Feature Detectors and Descriptors, vol. 630, Eds. Cham: Springer International Publishing, 2016, pp. 11–45.

[2] Y. M. Furgala and B. P. Rusyn, "Peculiarities of melin transform application to symbol recognition," 2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, pp. 251-254, 2018

[3] Y. Furgala, A. Velhosh, S. Velhosh and B. Rusyn, "Using Color Histograms for Shrunk Images Comparison," 2021 IEEE 12th International Conference on Electronics and Information Technologies (ELIT), Lviv, Ukraine, pp. 130-133, 2021

[4] D. G. Lowe, “Distinctive image features from scale-invariant keypoints” International Journal of Computer Vision, vol.60, issue 2, pp. 91-110, 2004.

[5] S. Leutenegger, M. Chli and R. Y. Siegwart, "BRISK: Binary Robust invariant scalable keypoints," 2011 International Conference on Computer Vision, Barcelona, Spain, pp. 2548-2555, 2011

[6] Ethan Rublee, Vincent Rabaud, Kurt Konolige, Gary Bradski. “ORB: an efficient alternative to SIFT or SURF”, 2011 IEEE International Conference on Computer Vision, pp.2564-2571, 2011.

[7] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool, “SURF: Speeded Up Robust Features”, Computer Vision and Image Understanding, vol. 110, № 3. – pp. 346-359, 2008.

[8] C.Michael, L. Vincent, S. Christoph, F.Pascal, “BRIEF: Binary Robust Independent Elementary Features”. CVLab, EPFL. – 2009

[9] P. M. Panchal, S. R. Panchal, and S. K. Shah, “A Comparison of SIFT and SURF,” International Journal of Innovative Research in Computer and Communication Engineering, vol. 1, no. 2, pp. 323-327, 2013

[10] E. P. Yudha, N. Suciati and C. Fatichah, "Preprocessing Analysis on Medical Image Retrieval Using One-to-one Matching of SURF Keypoints," 2021 5th International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, pp. 160-164, 2021

[11] A. Marmol, T. Peynot, A. Eriksson, A. Jaiprakash, J. Roberts and R. Crawford, "Evaluation of Keypoint Detectors and Descriptors in Arthroscopic Images for Feature-Based Matching Applications," in IEEE Robotics and Automation Letters, vol. 2, no. 4, pp. 2135-2142, Oct. 2017

[12] M. Ihmeida and H. Wei, "Image Registration Techniques and Applications: Comparative Study on Remote Sensing Imagery," 2021 14th International Conference on Developments in eSystems Engineering (DeSE), Sharjah, United Arab Emirates, 2021, pp. 142-148

[13] Y. Furgala, Y. Mochulsky and B. Rusyn, "Evaluation of Objects Recognition Efficiency on Mapes by Various Methods," 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, pp. 595-598, 2018

[14] Y. Li, J. Pan, L. Jiang and Y. Sun, "Based on Harris corner of Drones vehicle target image matching method," 2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, pp. 652-656, 2022

[15] R. Dijaya, N. Suciati and A. Saikhu, "Corn Plant Disease Identification Using SURF-based Bag of Visual Words Feature," 2022 14th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia, pp. 206-210, 2022

[16] G. Bradski and A. Kaehler, “Learning OpenCV: Computer vision with the OpenCV library.” O'Reilly Media, Inc., 2008

Identification of Internal Planar Square Defects in Composite Panels Using Optoacoustic Technique

Leonid Muravsky
Karpenko Physico-Mechanical Institute
of the NAS of Ukraine
Lviv, Ukraine
muravskyleon@gmail.com

Zinoviy Nazarchuk
Karpenko Physico-Mechanical Institute
of the NAS of Ukraine
Lviv, Ukraine
zinoviy.nazarchuk@gmail.com

Oleksandr Kuts
Karpenko Physico-Mechanical Institute
of the NAS of Ukraine
Lviv, Ukraine
kuts@ipm.lviv.ua

Oleksiy Sharabura
Karpenko Physico-Mechanical Institute
of the NAS of Ukraine
Lviv, Ukraine
shom@ipm.lviv.ua

***Abstract* — A new approach for determining the size and depth of internal square planar defects in laminated composite panels using a developed optoacoustic technique is proposed. The size of the defect is determined by spatial light responses generated by a region of interest, which is located above the defect and oscillated at the fundamental frequency or multiple resonant frequencies. The defect depth is determined using the family of experimental dependencies of the resonant frequency on the defect size. Experimental results proved the effectiveness of the proposed approach in identifying internal planar square defects in fiberglass panels.**

Keywords* — *internal square planar defects, composite panel, defect size and depth, optoacoustic technique, region of interest, dynamic speckle patterns, decorrelation

I. Introduction

Composite materials are widely used in machines, aircrafts, vehicles, buildings, etc. However, the presence of several components in composites with different physicochemical and mechanical characteristics contributes to the formation of various internal defects, including delaminations, cracks, disbonds, voids, inclusions, matrix cracking, etc. Therefore, the problem of detecting subsurface damage and defects in composite structures is very relevant [1].

Several nondestructive testing (NDT) techniques can detect internal defects in composites with a certain level of confidence [1, 2]. They also include hybrid optical-acoustic NDT techniques, in which the composite structure is excited by an acoustic wave, and its probing is carried out using optical radiation. These techniques can be divided conventionally on two basic directions. The first direction uses scanning laser beam to monitor the composite surface excited by acoustic waves [3–6], and second ones uses the expanded laser beam illumination of the excited surface area to record speckle fringe patterns, shearograms, holograms and speckle patterns [7–14].

We are developing the second direction based on Speckle Metrology techniques. In particular, in Karpenko Physico-Mechanical Institute of the NAS of Ukraine, the difference dynamic digital speckle pattern interferometry method for detection of internal defects in composite panels has been developed [11, 14]. The hybrid interferometric system that implements this method has been created and experiments to detect internal defects were performed. However, the interferometric systems implementing this and similar methods require complex equipment. In addition, these methods are very sensitive to vibrations and to speckle decorrelation.

II. Optoacoustic Technique

A new optoacoustic technique for detecting internal defects in composite structures allowing to reduce or even eliminate the influence of external vibrations has been developed [10, 11, 15, 16]. This technique is based on a new approach to defect detection by forming dynamic speckle patterns of the composite surface area excited by acoustic wave and selecting the region of interest (ROI) located directly above the defect. A spatial response from the defect is produced after digital processing of a series of dynamic speckle patterns. The hybrid optical-digital system (HODS) that implements this technique can be used in the field, since it is much less sensitive to external influences (vibrations, etc.) compared to known interferometric systems for detecting internal defects and is much simpler than such systems.

Upon acoustic excitation of the composite, the ROI above the defect begins to oscillate. ROI can be considered as a thin edge clamped membrane. Under the action of flexural acoustic waves, the ROI oscillates at its resonant frequencies, which approximately correspond to the resonant frequencies of such a membrane. Flexural waves act on the material in a direction transverse to the direction of acoustic wave propagation. Therefore, vibrations of the ROI in an out-of-plane direction concerning the surface are dominant. The simplified optical scheme of an imaging system for internal defects visualization is shown in Fig. 1. In this scheme, the composite panel containing the ROI is excited by acoustic flexural waves and the laser beam illuminates the studied rough surface area. The ROI begins to oscillate, and a series of dynamic speckle patterns of the illuminated area are recorded as the ROI oscillates. The vibrating ROI generates a local speckle pattern (LSP). In the LSP, its structure and speckle contrast change in sync with ROI oscillations. These changes are caused by the fact that the spatial frequency shift $\Delta\nu$ in the lens aperture plane due to the ROI's elements tilt on angle β leads to decorrelation of the LSP during the ROI tilting relative to the initial LSP,

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

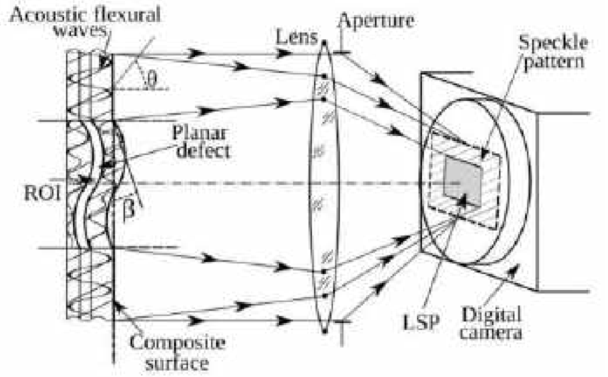


Fig. 1. Simplified optical scheme of the defect visualization.

when the ROI has no tilt. The angle β is connected with frequency shift Δν by the next equation [17]:

$$\Delta\nu=(1+\cos\theta)\beta/\lambda M_l, \quad (1)$$

where λ and θ are the wavelength and incidence angle of laser light, M_l is the lateral demagnification of the imaging system.

On the other hand, the LSP decorrelation can be estimated by the Yamaguchi correlation factor (YCF) C_Y, that is [18, 19]

$$C_Y = (4/\pi^2)\{\arccos(|\Delta\nu|/D_\nu) - (|\Delta\nu|/D_\nu)[1-(|\Delta\nu|/D_\nu)^2]^{0.5}\}^2, \quad (2)$$

where $D_\nu = D/\lambda z$ is the lens aperture diameter in the frequency domain, D is the lens aperture diameter, z is the distance from the lens aperture to the matrix sensor of a digital camera (DC) (see Fig. 1).

The YCF shows that full decorrelation is achieved at a given threshold tilt angle β_t if $\Delta\nu_t = D_\nu$. In this case, β_t is the minimum tilt angle at which the YCF between the transformed parts of the LSP generated by the ROI's tilted elements and the corresponding parts of the initial LSP generated by the flat ROI reaches zero. Taking into account (1) and (2), the threshold tilt angle is given by [15]

$$\beta_t = M_l D/z(1+\cos\theta). \quad (3)$$

ROI above the square planar defect can be considered as the thin edge clamped square membrane and the ROI out-of-plane motion can be evaluated by equation of the membrane motion [20, 21]. Flexural waves propagating in the ROI act in the perpendicular direction to the direction of propagation [3, 4, 22]. Therefore, only transverse nodes are subjected to tilting. Fig. 2 shows the transverse node tilts in a square membrane at the fundamental resonant frequency (membrane mode (1,1)) and multiple resonant frequency (membrane mode (1,2)). The membrane transverse nodes correspond to light spots. These spots are visualized due to the LSP decorrelation, which is performed by subtraction of the LSP generated by the tilted ROI from the LSP generated by the flat ROI.

The fundamental resonant frequency of the square planar defect is described by the next formula [4]:

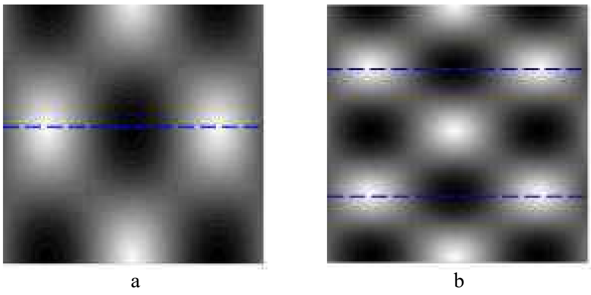


Fig. 2. Light spots in transverse nodes of an oscillated square membrane: (a) membrane mode (1,1); (b) membrane mode (1,2). Dashed lines show the direction of flexural waves action.

$$f_{11,t} = 1.71(h/a^2)[E/\rho(1-\mu^2)]^{0.5}, \quad (4)$$

where a is the defect size, h is the defect depth, E is the elastic modulus, ρ is the material density, and μ is the Poisson's ratio.

Knowing the fundamental frequency, it is easy to establish multiple resonant frequencies at mode (m,n) according to the given formula [23]

$$f_{mn,t} = (c/2)[(m^2+n^2)/a^2], \quad (5)$$

where c is the acoustic wave velocity.

Speckle blurring and decrease in speckle contrast in the LSP can also occur due to the reduction of spatial coherence of light during the ROI's out-of-plane displacement [24, 25], as well as due to the time averaging of dynamic speckles at their registration. This factor contributes to the formation of light spatial responses not only from the transverse nodes, but also from the antinodes of the oscillating ROI.

III. Implementation of Optoacoustic Technique

We have created the HODS setup implementing the developed optoacoustic technique [15, 16]. The setup scheme is shown in Fig. 3. In contrast to interferometric systems, it does not contain the reference beam, so it is not sensitive to vibrations and can be used in natural conditions. To extract the spatial response from a defect, we correlate the total speckle pattern $I_{q,o}(i,j)$ obtained at the maximum values of the acoustic wave amplitude with the total speckle pattern $I_{q,e}(i,j)$ obtained at its minimum values. These two speckle patterns are recorded at the opposite tilts of the ROI. The correlation procedure is performed by subtracting these patterns and obtaining a total difference speckle pattern

$$I_q(i,j)=|I_{q,o}(i,j) - I_{q,e}(i,j)|. \quad (6)$$

The total speckle patterns $I_{q,o}(i,j)$ and $I_{q,e}(i,j)$ are obtained by summing the initial speckle patterns $I_{p,q1}(i,j)$ and $I_{p,q2}(i,j)$, that is

$$I_{q,o}(i,j)=\Sigma_p[I_{p,q1}(i,j)], \quad I_{q,e}(i,j)=\Sigma_p[I_{p,q2}(i,j)]. \quad (7)$$

The initial speckle patterns are recorded during two exposures of the DC equal to frame time T. On the first frame time T, the series of initial speckle patterns $I_{p,q1}(i,j)$ are recorded at the maximum amplitude of the acoustic wave with a time gap τ<<T, which duration is controlled by an acousto-optic deflector. On the second frame time T, the series of initial speckle patterns $I_{p,q2}(i,j)$ are recorded at the

minimum amplitude of the acoustic wave with the same time gap τ.

In this setup, the full decorrelation between speckle patterns $I_{q,o}$ and $I_{q,e}$ is achieved with opposite tilts of the ROI. Hence, the threshold angle β_{thr} becomes two times smaller, that is, $\beta_{thr} = \beta_t/2$ (see (3)).

IV. Experiments on Identification of Square Planar Defects

We conducted experiments on the identification of internal square planar defects in fiberglass laminated panels made from layers of STEF-1 fiberglass and epoxy phenol polymer resin as a binding material. To this end, 6 test samples were produced. Each sample contains 3 panels glued together (400x250 mm). The bottom layer with a thickness of 5 mm is the base. The middle layer is 1.5 mm thick fiberglass, in which holes of various sizes and shapes were milled. The top panels are 0.5 to 3.0 mm thick in 0.5 mm increments. Arrangement of holes in the middle layer of panels is shown in Fig. 4.

Results of detection of spatial responses from the planar defect No.4 with sizes 20×20×1.5 mm^3 at the fundamental resonant frequency f_{11}=12.2 kHz and the depth h=0.41 mm and from the same defect at the multiple resonant frequency f_{12}=35.5 kHz and the depth h=1.87 mm are shown in Fig. 5. This Fig. shows that the structure of the spatial responses is similar to the structure of the transverse nodes of the square membrane (see Fig. 2), and the sizes of these responses are equal to the size of defect No.4.

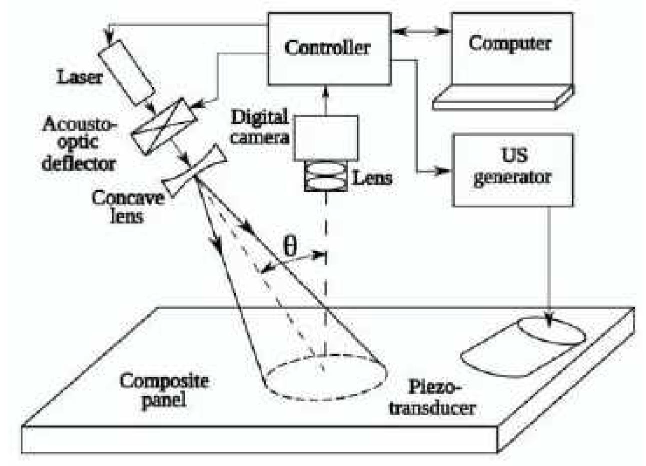


Fig. 3. Scheme of the hybrid optical-digital system setup.

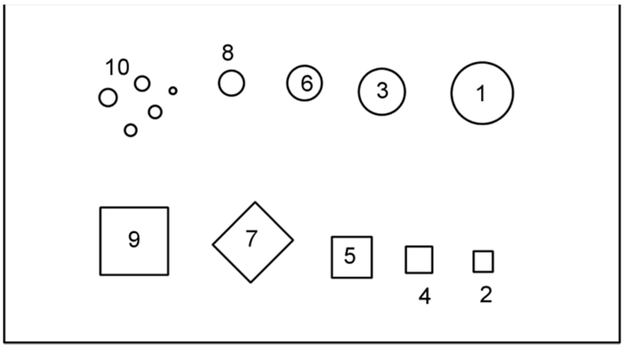


Fig. 4. Scheme of hole arrangement in middle layer of fiberglass laminated panels: 1 – round hole, ∅45 mm; 2 – square hole, 15×15 mm; 3 – round hole, ∅ 35 mm; 4 – square hole, 20×20 mm; 5 – square hole, 30×30 mm; 6 – round hole, ∅25 mm; 7 – square hole, 35×35 mm; 8 – round hole, ∅ 20 mm; 9 – square hole, 50×50 mm; 10 – round holes, ∅6, ∅8, ∅10, ∅12, and ∅14 mm.

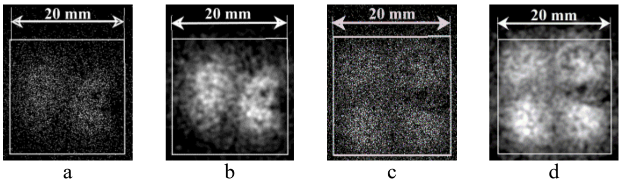


Fig. 5. Spatial responses from the defect No.4 at resonant frequencies: (a) response at the fundamental frequency f_{11}=12.2 kHz and the depth h=0.41 mm; (b) the same response after low-pass filtering (LPF); (c) response at the multiple frequency f_{12}=35.5 kHz and the depth h=1.87 mm; (d) the same response after LPF.

Structures of spatial responses from the square membrane at the multiple resonant frequencies f_{13} and f_{14} and results of detection of spatial responses from the same defect at the multiple resonant frequencies f_{13}=24.2 kHz and f_{14}=35.8 kHz and the depth h=0.41 mm are shown in Fig. 6. These Figs. also indicate that the structure of spatial responses is similar to the structure of the transverse nodes of the square membrane, and the dimensions of the obtained spatial responses correspond to the dimensions of the defect No.4.

We obtained dependencies of the experimental resonant frequencies f_{11}, f_{12}, f_{13} on the depth of the defect No.4 shown in Fig. 7 and marked by solid curves. The theoretical dependencies $f_{11,t}$, $f_{12,t}$, $f_{13,t}$ on the same defect depth were obtained using Eq. 4 and Eq. 5. To calculate these dependencies, it is necessary to know elastic modulus E, Poisson's ratio μ and the material density of ρ of the fiberglass panel STEF-1. To this end, the experiments for definition of E and μ were performed using the STEF-1 specimens, which thickness was equal to 1.8 mm. The orthotropic specimens were cut out along one of the two fiberglass laying directions, while the longitudinal axes of the specimens coincided with the direction of acoustic wave propagation. The fabricated specimens were fixed on a tensile-testing machine FP–100 and subjected to stretching until their rupture at F_{max}≈11,000÷11,800 N. An optoelectronic system containing a lens Jupiter-37A, a lens adapter and a digital camera BFS-U3-28S5 was used to measure the specimen tension $\Delta l_{||}$ in the longitudinal direction and compression $\Delta l_{\perp}$ in the transverse direction. The optical scheme of the system was designed so that its linear magnification was equal to M=0.3. Based on experimental data and results of calculations, we established that E=23.3±1.3 Gpa, μ=0.14±0.02, ρ=(1.70±0.05)x10^3 kg/m^3. These experimental data allow calculating the theoretical fundamental and multiple resonant frequencies. In Fig. 7, the theoretical resonant frequencies $f_{11,t}$, $f_{12,t}$, $f_{13,t}$ are marked by dashed lines. All of them indicate a monotonic increase in these frequencies with increasing the defect depth h. The dependences of the experimental and theoretical resonant frequency ratios on the defect depth h are shown in Fig. 8. As we can see, the deviations of the experimental results from the theoretical ones are clearly seen in both graphs. The identified discrepancies between these dependencies can be explained, in particular, by imperfection of formula (4) for determining the fundamental resonant frequency $f_{11,t}$, deviations in the defects size, the material orthotropy, and deviations in the direction of acoustic wave propagation. Similar experiments were performed with the square planar defects 2, 5, 7 and 9 shown in Fig. 4.

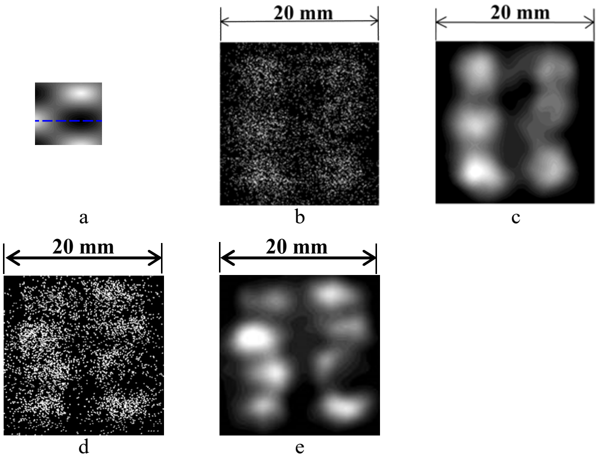


Fig. 6. Spatial responses from the defect No.4 at multiple resonant frequencies: (a) structure of square membrane nodes at the resonant frequency f_{13}, where transverse nodes are intersect with dashed lines; (b) response at the resonant frequency f_{13}=24.2 kHz and the depth h=0.41 mm; (c) the same response after low-pass filtering (LPF); (d) response at the resonant frequency f_{14}=35.8 kHz and the depth h=0.41 mm; (e) the same response after LPF.

The obtained graphic dependencies in Fig. 7 and Fig. 8 indicate that, knowing the spatial responses sizes and the fundamental resonant frequencies f_{11} values, it is possible to determine the depth h of each found planar square defect. To this end, the family of experimental dependencies of the fundamental resonant frequencies on the size of defects, shown in Fig. 9 can be used to find the depth of the detected planar defect. One theoretical dependence for resonant frequencies $f_{11,t}$ and the depth h=0.41 mm is shown in this Fig. as a dashed curve. The reasons for the discrepancies between the experimental and theoretical dependences are indicated above. A particularly strong discrepancy takes place for deep and short defects caused by the deviation of the vibration mode from the classical approximation of a thin plate. To reduce the error in determining the depth of defects, families of dependencies of multiple resonant frequencies f_{12}, f_{13}, f_{14}, f_{15}, etc. on the defects size can also be built. An example of the family of dependencies of the resonant frequency f_{12} on the defect size is shown in Fig. 10.

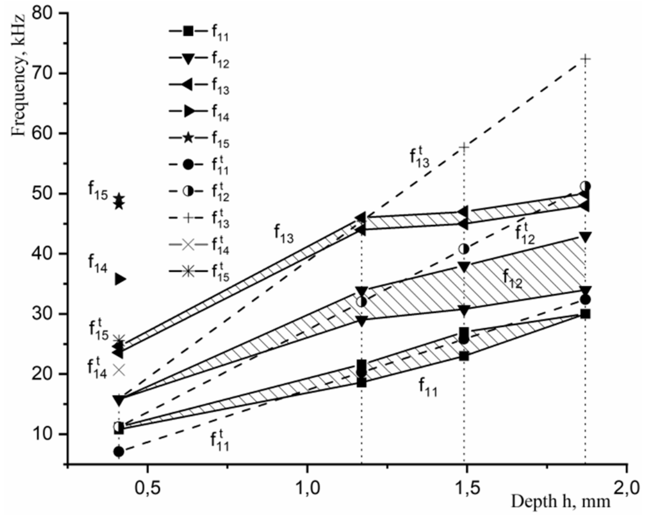


Fig. 7. Dependencies of experimental and theoretical resonant frequencies f_{11}, f_{12}, f_{13} and f^t_{11}, f^t_{12}, f^t_{13}, (dashed lines) on the depth of the defect No. 4. Frequencies f_{11}: ● – ⬟; f_{12}: ▼ – ▲; f_{13}: ◀ – ◆; f_{14}: ▶; f_{15}: ★ – ⬢.

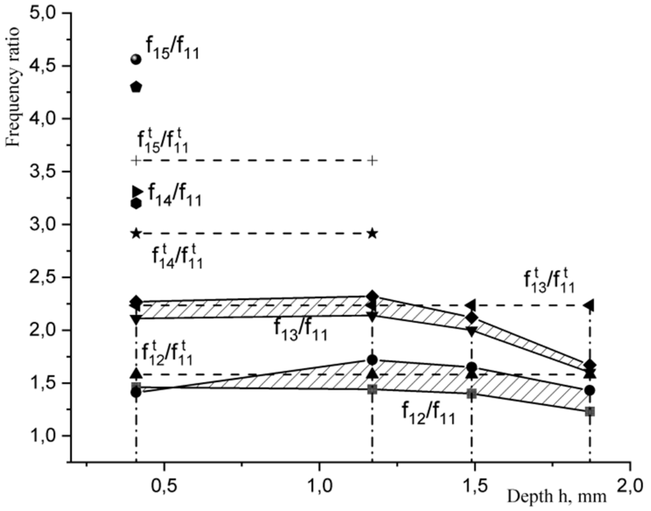


Fig. 8. Dependencies of experimental and theoretical frequency ratios f_{12}/f_{11}, $f_{13},/f_{11}$ and f^t_{12}/f^t_{11}, f^t_{13}/f^t_{11} (dashed lines) on the depth of the defect No. 4. Frequency ratios: f_{12}/f_{11}:● – ■; f_{13}/f_{11}:◆ – ▼; f_{14}/f_{11}:▶ – ⬢; f_{15}/f_{11}:◉ – ⬟; f^t_{12}/f^t_{11}:▲; f^t_{13}/f^t_{11}:◀ ; f^t_{14}/f^t_{11}:★; f^t_{15}/f^t_{11}: +.

Conclusions

Thus, the new approach to determine the size and depth of square planar subsurface defects in composite panels using the spatial structure is proposed using the NDT optoacoustic technique. Internal square planar defects can be detected in fiberglass laminated composite panels at the depths h=0.1÷3 mm using the developed optoacoustic technique. These defects can also be detected at the deeper depths. The fundamental or multiple resonant frequencies of the planar square defect excitation can be determined from the spatial structure of the defect light responses within the ROI. Formation of light responses from defect occurs due to the tilting of the ROI under acoustic excitation. The size of the light response is approximately equal to the defect size. Hence, knowing the defect size and the fundamental resonant frequency of the ROI, we can find the defect depths using the

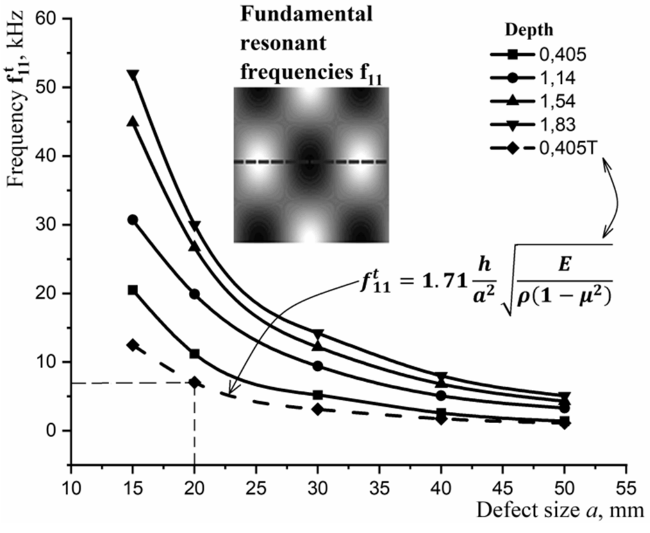


Fig. 9. Dependencies of fundamental resonant frequencies f_{11} (solid curves) and f^t_{11} (dashed curve) on the defects size a for different depths.

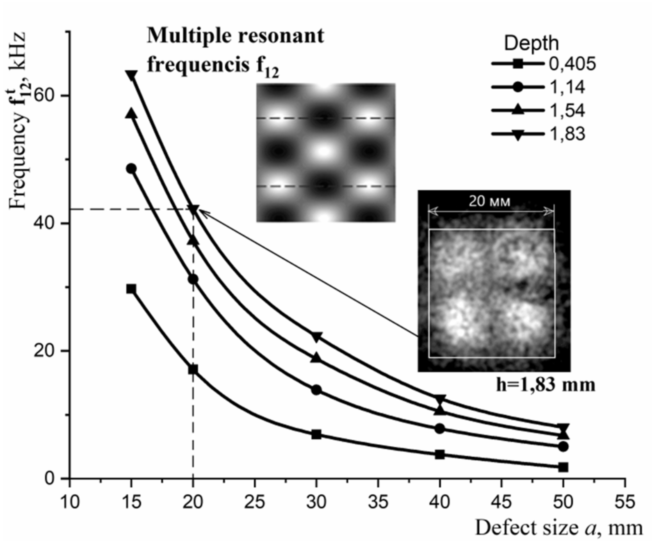


Fig. 10. Dependencies of multiple resonant frequencies f_{12} on the defects size a for different depths.

dependencies of the fundamental resonant frequency on the defect size (see Fig. 9). To improve the accuracy of the defect depth definition, one can use the dependencies of the multiple resonant frequencies on the defect size (see, for example, Fig. 10).

References

[1] V. M. Karbhari, Ed., Non-Destructive Evaluation (NDE) of Polymer Matrix Composites. Oxford, Cambridge: Woodhead Publishing Lim., 2013.

[2] S. Sfarra, N. P. Avdelidis, C. Ibarra-Castanedo, C. Santulli, P. Theodorakeas, A. Bendada, D. Paoletti, M. Koui, and X. Maldague, “Surface and subsurface defects detection in impacted composite materials made by natural fibers, using nondestructive testing methods,” International Journal of Composite Materials, vol. 4, no. 5A, pp. 1-9, 2014.

[3] I . Solodov, J. Bai, and G. Busse, “Resonant ultrasound spectroscopy of defects: case study of flat-bottomed holes,” Journal of Applied Physics, vol. 113, no. 22, pp. 223512, 2013.

[4] I. Solodov, M. Rahammer, and M. Kreutzbruck, “Analytical evaluation of resonance frequencies for planar defects: Effect of a defect shape,” NDT & E International, vol. 102, pp. 274-280, 2019.

[5] J. Segers, S. Hedayatrasa, G. Poelman, W. Van Paepegem, and M. Kersemans, “Probing the limits of full-field linear local defect resonance identification for deep defect detection,” Ultrasonics, vol. 105, pp. 106130, 2020.

[6] J. Segers, S. Hedayatrasa, G. Poelman, W. Van Paepegem, and M. Kersemans, “Robust and baseline-free full-field defect detection in complex composite parts through weighted broadband energy mapping of mode-removed guided waves,” Mechanical Systems and Signal Processing, vol. 151, pp. 107360, 2021.

[7] L. S. Wang and S. Krishnaswamy, “Additive-subtractive speckle interferometry: extraction of phase data in noisy environments,” Optical Engineering, vol. 35, no. 3, pp. 794-801, 1996.

[8] P. Fomitchov, L. S. Wang, and S. Krishnaswamy, “Advanced image-processing techniques for automatic nondestructive evaluation of adhesively-bonded structures using speckle interferometry,” Journal of Nondestructive Evaluation, vol. 16, pp. 215-227, 1997.

[9] Á. F. Doval, C. Trillo, D. Cernadas, B. V. Dorrío, C. López, J. L. Fernández, and M. Pérez-Amor, “Measuring amplitude and phase of vibration with double-exposure stroboscopic TV Holography.” in Interferometry in Speckle Light: Theory and Applications. Berlin, Heidelberg: Springer, 2000, pp. 281-288.

[10] L. Muravsky, O. Kuts, G. Gaskevych, and O. Suriadova, “Detection of subsurface defects in composite panels using dynamic speckle patterns,” in Proc. IEEE XIth International Scientific and Practical Conference on Electronics and Information Technologies ’09, 2019, pp. 7-10.

[11] Z. Nazarchuk, L. Muravsky, and D. Kuryliak, “To the problem of the subsurface defects detection: theory and experiment,” Procedia Structural Integrity, vol. 16, pp. 11-18, 2019.

[12] B. P. Thomas, S. A. Pillai, and C. S. Narayanamurthy, “Investigation on vibration excitation of debonded sandwich structures using time-average digital holography,” Applied Optics, vol. 56, no. 13, pp. F7-F13, 2017.

[13] B. P. Thomas, S. A. Pillai, and C. S. Narayanamurthy, “Computed time average digital holographic fringe pattern under random excitation,” Applied Optics, vol. 60, no. 4, pp. A188-A194, 2021.

[14] Z. Nazarchuk, L. Muravsky, and D. Kuryliak, “Digital speckle pattern interferometry for studying surface deformation and fracture of materials,”. in Optical Metrology and Optoacoustics in Nondestructive Evaluation of Materials. Springer Series in Optical Sciences, vol. 242. Singapore: Springer, 2023, pp. 149-217.

[15] Z. T. Nazarchuk, L. I. Muravsky, and O. G. Kuts, “Nondestructive testing of thin composite structures for subsurface defects detection using dynamic laser speckles,” Research in Nondestructive Evaluation, vol. 33, no. 2, pp. 59-77, 2022.

[16] Z. Nazarchuk, L. Muravsky, and D. Kuryliak, “Methods for processing and analyzing the speckle patterns of materials surfaces,” in Optical Metrology and Optoacoustics in Nondestructive Evaluation of Materials. Springer Series in Optical Sciences, vol. 242. Singapore: Springer, 2023, pp. 249-323.

[17] T. Fricke-Begemann, “Three-dimensional deformation field measurement with digital speckle correlation,” Applied Optics, vol. 42, no. 34, pp. 6783-6796, 2003.

[18] I. Yamaguchi, “Speckle displacement and decorrelation in the diffraction and image fields for small object deformation,” Optica Acta: International Journal of Optics, vol. 28, no. 10, pp. 1359-1376, 1981.

[19] I. Yamaguchi, “Theory and applications of speckle displacement and decorrelation,” in Speckle Metrology, R. C. Sirohi, Ed., Boca Raton, FL: CRC Press, 2020, pp. 1-40.

[20] I. G. Aramanovich and V. I. Levin, Equations of mathematical physics. M.: Nauka, 1969.

[21] L. Keene and F. P. Chiang, “Real-time anti-node visualization of vibrating distributed systems in noisy environments using defocused laser speckle contrast analysis,” Journal of Sound and Vibration, vol. 320, no. 3, pp. 472-481, 2009.

[22] F. Ciampa, S. G. Pickering, G. Scarselli, and M. Meo, “Nonlinear imaging of damage in composite structures using sparse ultrasonic sensor arrays,” Structural Control and Health Monitoring, vol. 24, no. 5, pp. e1911, 2017.

[23] D. A. Russell, Graduate Program in Acoustics, The Pennsylvania State University. Topic: “Vibrational Modeshapes of a Rectangular Membrane (fixed at the edges).” Available: https://www.acs.psu.edu/drussell/Demos/rect-membrane/rect-mem.html [Completely rewritten on October 18, 2018].

[24] B. Eliasson and F.M. Mottier, “Determination of the granular radiance distribution of a diffuser and its use for vibration analysis,” Journal of the Optical Society of America, vol. 61, no. 5, pp. 559-565, 1971.

[25] M. Owner-Petersen, “Decorrelation and fringe visibility: on the limiting behavior of various electronic speckle-pattern correlation interferometers,” Journal of the Optical Society of America A, vol. 8, no. 7, pp. 1082-1089, 1991.

Bluetooth Low-Energy Beacon Resistance to Jamming Attack

Volodymyr Sokolov
Dept. of Inform. and Cyber Security
Borys Grinchenko Kyiv University
Kyiv, Ukraine
0000-0002-9349-7946

Pavlo Skladannyi
Dept. of Inform. and Cyber Security
Borys Grinchenko Kyiv University
Kyiv, Ukraine
0000-0002-7775-6039

Volodymyr Astapenya
Dept. of Inform. and Cyber Security
Borys Grinchenko Kyiv University
Kyiv, Ukraine
0000-0003-0124-216X

***Abstract* — To assess the potential damage from jamming attacks on a Bluetooth Low-Energy (BLE) Beacon device, an experimental setup with a packet emitter, sniffer, and signal spectrum control at the receiving point was used. With less than 0.7% of successfully delivered packets, it is impossible to talk about the success of even non-critical equipment. From the ratio of single-error packets to lost packets rate, it can be seen that the tipping point occurs on the chart at minus 12 dBm for a 2 Mb/s baud rate. An increase in the level of the interference signal may be invisible to the user until a certain point when the number of errors in the system begins to increase like an avalanche. This results in an increase in lost packets and the generation of new traffic due to retransmissions. However the additional traffic is also subject to interference, so the useful data rate is reduced even more. It should be noted that for low transmission rates, the single-error packets to lost packets rate is constant.**

***Keywords* — jamming, Bluetooth Low-Energy, BLE, beacon, iBeacon, Eddystone, AltBeacon, PER.**

I. Introduction

The widespread use of passive radio tags in retail has simplified the process of selling goods. The cost of the radio tag has little effect on the price of the goods, therefore it can be widely used for goods of the middle and highest price categories. For advertising, information, and navigation purposes inside the trading floor, tags of another type of BLE Beacon are used, which actively transmit their identifiers. Simultaneously with labels, local and global networks are used, which allows us to process actions in real time [1–3].

Jamming is an attack that disrupts the entire wireless network and restricts all or part of the availability of resources on a given network [4].

In previous works, we have considered various types of attacks on wireless systems: sniffing [5], spoofing [5, 6], denial-of-service [7], and men-in-the-middle attacks [8], including using Software-Defined Radio (SDR) [5, 8]. The goal of the current study is to determine the resistance of BLE Beacon technology to jamming attacks.

Sect. II reviewed previous work in this area. In Sect. III presents the scheme of the experiment, hardware, and software. Victim signal level validation is described in Sect. IV. The results of the experiment are given in Sect. V (spectrum estimation) and VI (error rate). The paper ends with conclusions and outlines for the following research in Sect. VII.

II. Source Review

The proposed in [4] considers models of the node's round trip time to detect jamming attacks with the optimization method. This approach can only partially secure the system but does not completely solve the problem of the vulnerability of the wireless network. [9] provides an in-depth overview of various secure relaying strategies and schemes of solutions for cooperative jamming techniques, with an emphasis on power allocation and beamforming techniques. This distribution avoids a complete failure of the system. Individual elements may become inaccessible, so duplicate paths should be used to ensure accessibility. The adaptation method proposed in [10] requires additional systems for analysis and reconfiguration, which complicates the implementation of such trips. The Bayes-adaptive mixed-observable Markov decision process model is proposed in [11] to extend the current Bayesian reinforcement learning models to handle the state's mixed observability. This approach dramatically increases security but requires additional computing resources.

III. Experimental Layout

A. Experiment General Scheme

In the current research, a scheme was used in which BTE Beacon devices as victims. The active attacker is implemented on Nordic Semiconductor (NS) nRF24L01+ with power and low-noise amplifiers [12]. The packet sniffer is configured to receive packets from Beacons. The sniffer is based on Texas Instruments (TI) CC2541 [13]. A spectrum analyzer based on the SDR bladeRf [14] was used as a control device. The interaction scheme is shown in Fig. 1.

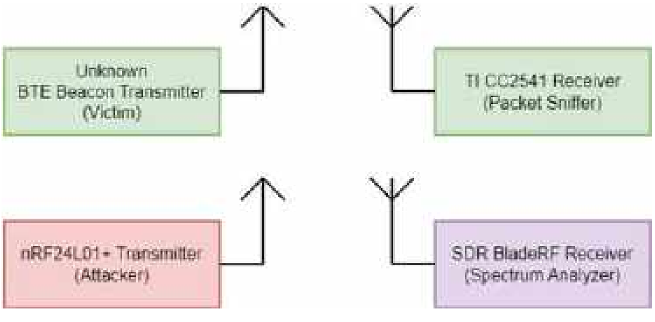


Fig. 1. Experiment scheme.

B. Victim Equipment

Modern Beacons work on BLE technology—Bluetooth 5.0 and higher. The most common devices are presented in Table I. The formats of transmitted messages differ [15, 16],

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

but at the moment devices are already available that can work with different formats simultaneously [17].

TABLE I. MOST COMMON BEACON TYPES

Beacon Type	Manufacture	Package length, bytes
iBeacon [18]	Apple Inc.	30
Eddystone [19]	Google Inc.	31
AltBeacon [20]	Radius Networks Inc.	37

The equipment operates in the same frequency range and according to the same specification, so we can use the same jamming method for all devices of this type.

C. Packet Sniffer Equipment

The TI CC2541 module [13] was used as a packet sniffer. Two module modifications of this sniffer are available, shown in Fig. 2.

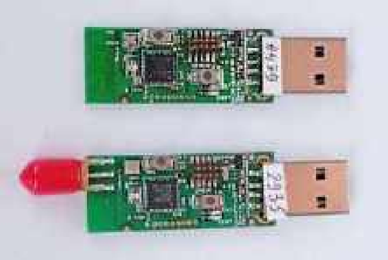


Fig. 2. Various designs of TI CC2541 packet sniffer with planar (above) and external antenna (below).

Installing the sniffing firmware requires the use of the TI CC-Debugger (see Fig. 3), which in turn requires a native firmware update [21].

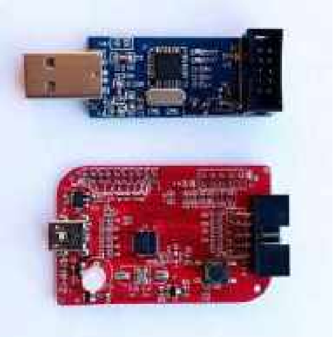


Fig. 3. Various designs of TI CC-Debugger replica (above) and original design (below).

The connection diagram of the TI CC2541 modules for flashing is shown in Fig. 4.

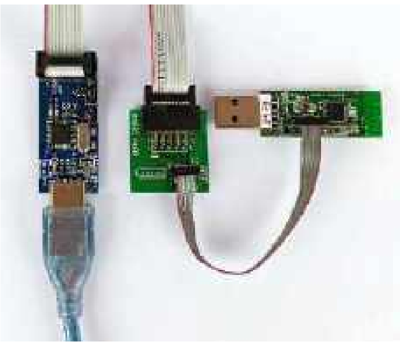


Fig. 4. Packet sniffer firmware scheme.

Sniffing results can be saved as a sequence of packets that are available for later analysis in the SmartRF Packet Sniffer software [22].

D. Attacker Equipment

The NS nRF24L01+ module [12] is used as an attacking device, which operates in a wider range than is required for jamming BLE. In addition, you do not need to jam all forty channels that the specification suggests, but only three channels: 37th (2.402GHz), 38th (2.426GHz), and 39th (2.480GHz). Two module modifications of this attacker are available, shown in Fig. 5.

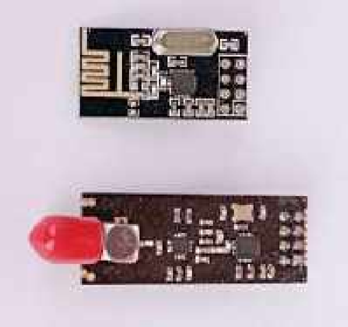


Fig. 5. Various designs of NS nRF24L01+ with planar antenna (above) and power amplifier, low-noise amplifier, and external antenna (below).

To implement the attacking side, the project described in [23] was used. The Arduino Nano module is used to control the NS nRF24L01+ module, as shown in Fig. 6. The attacker can work stand-alone with the external power supply. External control is not required for it.

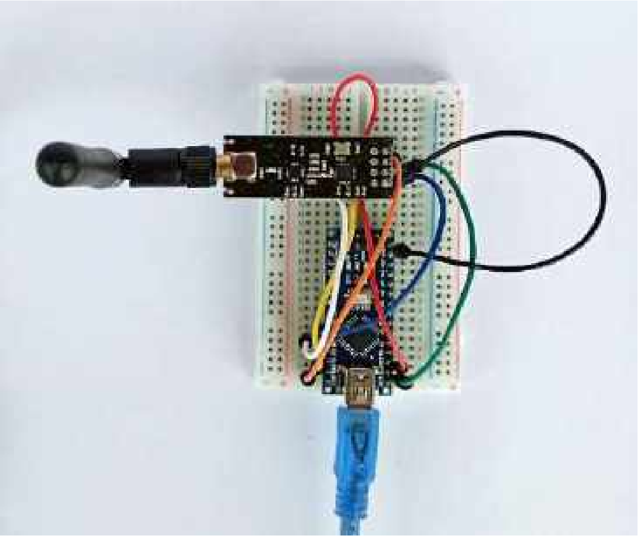


Fig. 6. General view of the attack module.

Unlike the proposed project, only one transmitting module is used. Before each packet is sent, the module must be reinitialized. As a result, the number of packets in the channel decreased by more than six times due to the time for re-initialization. Only the 38th (0×26) channel was used in the experiment (Fig. 7).

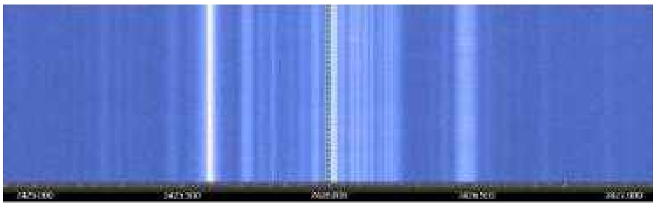


Fig. 7. An example of a waterfall diagram for the 38th channel.

E. Spectrum Analyzer Equipment

The multifunctional SDR Nuand bladeRF [14, 24] was used as a spectrum analyzer shown in Fig. 8. To obtain spectrograms, the SDR Console (ver. 3.2, build 2731) [25] with a driver for this type of SDR [26] is used. The data from the spectrum analyzer is corrective but is not used to characterize the success of an attack.

Fig. 8. SDR bladeRF as a spectrum analyzer.

F. Experimental Setup

The experiment was carried out inside the building. The transmitter and receivers were located at a height of one wavelength (~12.5 cm) from the floor and in the far field, which is a minimum of 60 cm (Fig. 9).

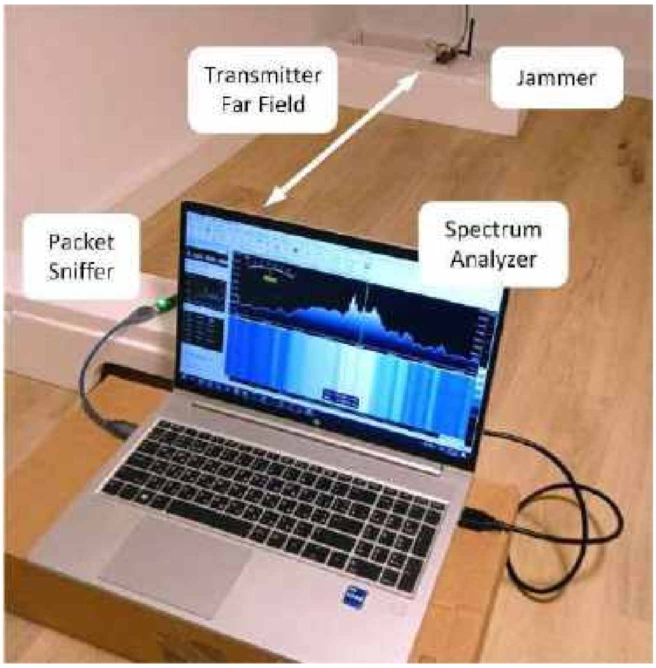
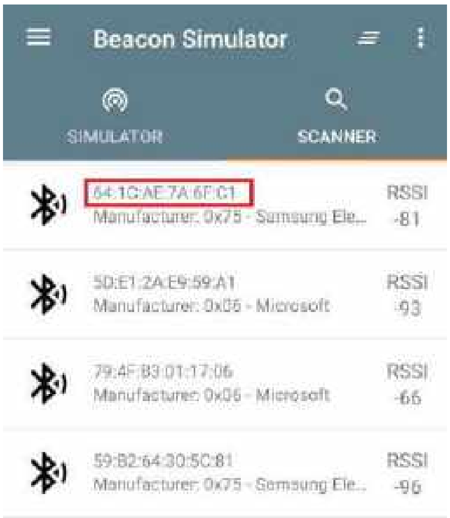


Fig. 9. General view of the experimental setup.

G. Checking for "Victims"

Using the Beacon Simulator application for Android phones [27], four beacon devices are identified in range (see Fig. 10).



64:1C:AE:7A:6F:C1
Manufacturer: 0x75 - Samsung Ele... RSSI -81

5D:E1:2A:E9:59:A1
Manufacturer: 0x06 - Microsoft RSSI -93

79:4F:B3:01:17:06
Manufacturer: 0x06 - Microsoft RSSI -66

59:B2:64:30:5C:81
Manufacturer: 0x75 - Samsung Ele... RSSI -96

Fig. 10. List of "victims" with identifiers.

The same packages are received in the SmartRF Packet Sniffer software [20] and shown in Fig. 11.

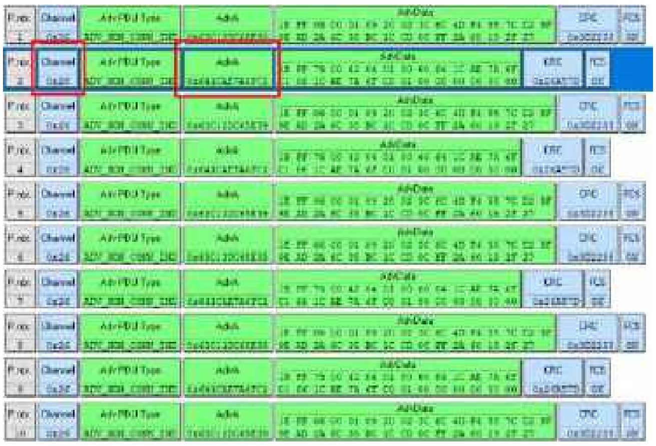


Fig. 11. Successfully received packages.

IV. Victim Signal Level Validation

To estimate the limits of the working signal level of the victims, we use the loss path model:

$$P_{RX} = P_{TX} - 10\gamma\log_{10}(r/r_0) - L_0 \qquad (1)$$

where P_{TX} is transmit power, dBm; γ is the path loss exponent; r is the length of the path; r_0 is the reference distance, $r_0 = 1$ m for microcell; L_0 is a normal random variable [28].

For our experiment, let's take the average value of path loss exponent $\gamma = 5$ (for indoor it is selected in the range from 4 to 6) and is the path length $r = 10$ m.

Maximum transmitter signal level $P_{TX\,max}$ = 4 dBm and receiver $P_{RX\,max}$ = –66 dBm (see Fig. 10) [29] therefore L_0 = 10 dBm. Thus, the devices are within the range of minus 66 to minus 100 dBm.

V. Spectrum Comparison

The signal spectrum was tracked to control and maintain the invariability of the experimental conditions. The instantaneous spectrum of the signal at the receiving point was obtained for the minimum (Fig. 12), medium (Fig. 13), and maximum (Fig. 14) transmitter power.

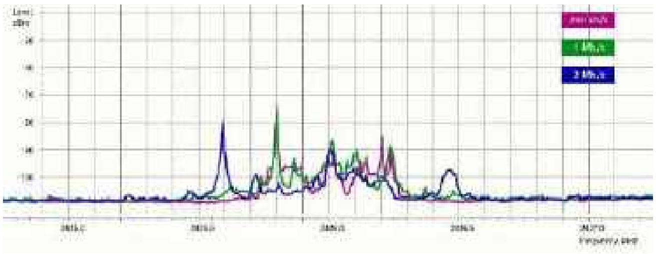


Fig. 12. Spectrum transmits power $P_{TX} = -36$ dBm for different baud rates.

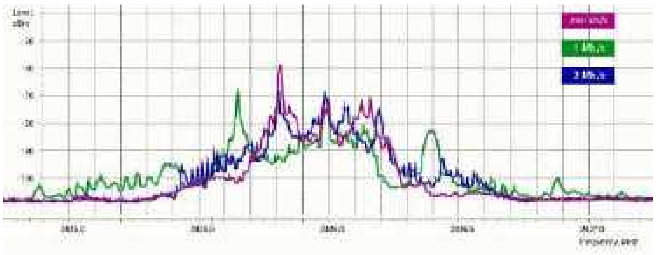


Fig. 13. Spectrum transmits power $P_{TX} = -18$ dBm for different baud rates.

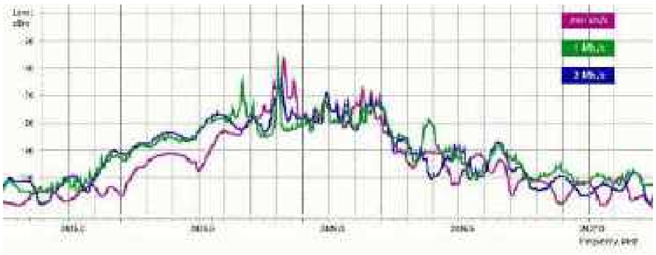


Fig. 14. Spectrum transmits power $P_{TX} = 0$ dBm for different baud rates.

It can be seen from the graphs that the larger the baud rate, the wider the spectrum becomes. This, in turn, leads to a lower concentration of energy in the BLE channel, and hence a decrease in the number of errors.

VI. Packet Error Rate

The experiments were carried out for different signal levels (0, –6, –12, –18, and –36 dBm) and different baud rates (250 kb/s, 1 Mb/s, and 2 Mb/s). For each combination, the experiment was repeated five times. 960 packets were sent, and as a result, the number of received and lost packets was analyzed. The received packets were of two types: integer and with a single error (CRC allows us to identify such packets). To evaluate the results, the Packet Error Rate (PER) is used as the ratio of whole packets to the total number of packets sent. The pattern is almost linear, as shown in Fig. 15.

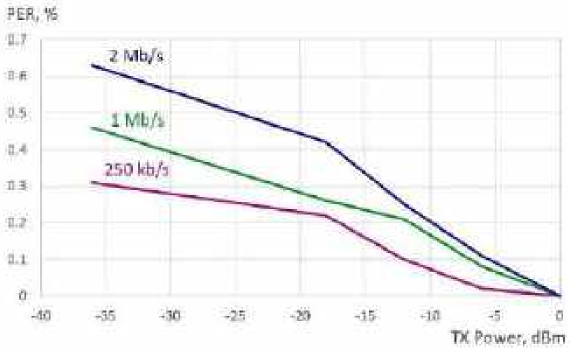


Fig. 15. PER depends on transmitter power and baud rate.

The turning point of the graph occurs for different data transfer rates in the range of minus (20..10) dBm. The smaller the transmitted packet, the later this fracture occurs. The process is intermittent, as single-error packets can be detected and corrected.

In addition, we can build a pattern of packets with a single error to lost packets (Fig. 16).

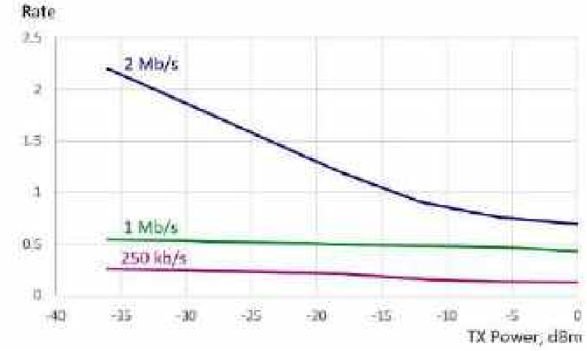


Fig. 16. Single-error packets to lost packets rate.

With a decrease in the transmission speed, no change in the graph is observed; only for a transmission speed of 2 Mb/s, with an increase in the interference power of more than minus 12 dBm single-error packets to lost packets rate becomes almost constant.

As can be seen from the above graphs, jamming allows us to almost completely drown out the transmission of information packets. Since ZigBee protocol is used to transmit data from sensors, one of the ways to ensure the stability of the network may be duplication at different frequencies or the use of a heterogeneous wireless network with proprietary data transfer protocols [30].

VII. Conclusions and Future Work

With less than 0.7% of successfully delivered packets, it is impossible to talk about the success of even non-critical equipment. From the ratio of single-error packets to lost packets rate, it can be seen that the tipping point occurs on the chart at minus 12 dBm for a 2 Mb/s baud rate. An increase in the level of the interference signal may be invisible to the user until a certain point when the number of errors in the system begins to increase like an avalanche. This results in an increase in lost packets and the generation of new traffic due to retransmissions. However the additional traffic is also subject to interference, so the useful data rate is reduced even more. It should be noted that for low transmission rates, the single-error packets to lost packets rate is constant.

In future studies, it is planned to consider the impact of jamming on the operation of commercial ZigBee modules.

References

[1] I. Kuzminykh, et al., “Investigation of the IoT Device Lifetime with Secure Data Transmission,” Lecture Notes in Computer Science. Springer International Publishing, pp. 16–27, 2019. doi: 10.1007/978-3-030-30859-9_2.

[2] F. Kipchuk, et al., “Investigation of Availability of Wireless Access Points based on Embedded Systems,” 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology. IEEE, 2019. doi: 10.1109/picst47496.2019.9061551.

[3] A. Carlsson, et al., “Sustainability Research of the Secure Wireless Communication System with Channel Reservation,” 2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics,

Telecommunications and Computer Engineering (TCSET). IEEE, 2020. doi: 10.1109/tcset49122.2020.235583.

[4] S. Murugaveni and B. Priyalakshmi, "Layering of Edge Node for Jamming Attack Detection and Elimination in Wireless Sensor Network," Concurrency and Computation: Practice and Experience. Wiley, May 2023. doi: 10.1002/cpe.7737.

[5] M. TajDini, V. Sokolov, and P. Skladannyi, "Performing Sniffing and Spoofing Attack Against ADS-B and Mode S using Software Define Radio," 2021 IEEE International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo). IEEE, pp. 7–11, 2021. doi: 10.1109/ukrmico52950.2021.9716665.

[6] Y. Sadykov, V. Sokolov, and P. Skladannyi, "Technology of Location Hiding by Spoofing the Mobile Operator IP Address," 2021 IEEE International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo). IEEE, pp. 22–25, 2021. doi: 10.1109/ukrmico52950.2021.9716700.

[7] V. Buryachok and V. Sokolov, "Using 2.4 GHz Wireless Botnets to Implement Denial-of-Service Attacks," Web of Scholar, vol. 6(24), no. 1, pp. 14–21, 2018. doi: 10.31435/rsglobal_wos/12062018/5734.

[8] M. TajDini, V. Sokolov, and V. Buriachok, "Men-in-the-Middle Attack Simulation on Low Energy Wireless Devices using Software Define Radio," 8th International Conference on "Mathematics. Information Technologies. Education:" Modern Machine Learning Technologies and Data Science, vol. 2386, pp. 287–296, 2019.

[9] F. Jameel, et al., "A Comprehensive Survey on Cooperative Relaying and Jamming Strategies for Physical Layer Security," IEEE Communications Surveys &amp; Tutorials, vol. 21, no. 3. Institute of Electrical and Electronics Engineers, pp. 2734–2771, 2019. doi: 10.1109/comst.2018.2865607.

[10] T. Erpek, Y. E. Sagduyu, and Y. Shi, "Deep Learning for Launching and Mitigating Wireless Jamming Attacks," IEEE Transactions on Cognitive Communications and Networking, vol. 5, no. 1. Institute of Electrical and Electronics Engineers, pp. 2–14, 2019. doi: 10.1109/tccn.2018.2884910.

[11] A. N. Elbattrawy, et al., "Model-based Bayesian reinforcement learning for enhancing primary user performance under jamming attack," Ad Hoc Networks, vol. 148. Elsevier BV, p. 103206, 2023. doi: 10.1016/j.adhoc.2023.103206.

[12] H.-T. Chen, P.-Y. Lin, and C.-Y. Lin, "A Smart Roadside Parking System Using Bluetooth Low Energy Beacons," Advances in Intelligent Systems and Computing. Springer International Publishing, pp. 471–480, 2019. doi: 10.1007/978-3-030-15035-8_44.

[13] D. Hernández-Rojas, et al., "Design and Practical Evaluation of a Family of Lightweight Protocols for Heterogeneous Sensing through BLE Beacons in IoT Telemetry Applications," Sensors, vol. 18, no. 2. MDPI AG, p. 57, Dec. 27, 2017. doi: 10.3390/s18010057.

[14] Feasycom, "FeasyBeacon 5Mart FSC-BP104 Bluetooth 5.0 Battery Powered Beacon Datasheet," Ver. 1.4. https://www.feasycom.net/Content/upload/pdf/201913049/FSC-BP104.pdf

[15] Apple, "Proximity Beacon Specification," Release R1, 2015. https://developer.apple.com/ibeacon/

[16] Google, "Eddystone," 2016. https://github.com/google/eddystone/tree/master/eddystone-uid

[17] AltBeacon, "Spec," 2019. https://github.com/AltBeacon/spec

[18] Texas Instruments, "2.4-GHz Bluetooth Low Energy and Proprietary System-on-Chip," 2013 https://www.ti.com/lit/ds/symlink/cc2541.pdf

[19] Texas Instruments, "CC-Debugger. Debugger and Programmer for RF System-on-Chips," 2019. https://www.ti.com/tool/CC-DEBUGGER

[20] Texas Instruments, "SmartRF Packet Sniffer. User's Manual," 2014. https://www.ti.com/lit/ug/swru187g/swru187g.pdf

[21] Nordic Semiconductor, "nRF24L01+. Single Chip 2.4 GHz Tranceiver. Preliminary Product Specification," ver. 1.0, 2008 https://www.sparkfun.com/datasheets/Components/SMD/nRF24L01Pluss_Preliminary_Product_Specification_v1_0.pdf

[22] W. Ler, "BLE-jammer," 2020. https://github.com/lws803/BLE-jammer

[23] J. Szymaniak, "BladeRF Windows Install Guide. Installing BladeRF Software with MatLab and Simulink Support," 2016. https://www.nuand.com/bladeRF-doc/guides/bladeRF_windows_installer.html

[24] "Is Bluetooth Low Energy Jamming Possible with an SDR Like the HackRF on GNURadio?" 2020. https://dsp.stackexchange.com/questions/70989/is-bluetooth-low-energy-jamming-possible-with-an-sdr-like-the-hackrf-on-gnuradio

[25] S. Brown, "V3.2 Release Notes," 2022. https://forum.sdr-radio.com:4499/viewtopic.php?f=66&t=1722

[26] J.-M. Picod, "SDRSharp-BladeRF," 2019. https://github.com/jmichelp/sdrsharp-bladerf

[27] V. Hiribarren, "Generate Nearby Notifications using Beacon Simulator," 2017. https://workshop.alea.net/post/2017/10/nearby-notifications-simulator/

[28] J. Klinglmayr, et al., "Sustainable Consumerism via Context-Aware Shopping," International Journal of Distributed Systems and Technologies, vol. 8, no. 4. IGI Global, pp. 54–72, Oct. 2017. doi: 10.4018/ijdst.2017100104.

[29] Kontakt, "Beacon Transmission Power, Range, and RSSI," 2020. https://support.kontakt.io/hc/en-gb/articles/4413258518930-Beacon-transmission-power-range-and-RSSI

[30] V. Sokolov, et al., "Method for Increasing the Various Sources Data Consistency for IoT Sensors," 2022 IEEE 9th International Conference on Problems of Infocommunications, Science and Technology (PICST). Oct. 10, 2022. doi: 10.1109/picst57299.2022.10238518.

Measuring Complex for Determination of Temperature Characteristics of Thermistors

Oleksandr Andreiev
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
Oleksandr.Andreiev@khpi.edu.ua

Olga Andreieva
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
Olga.Andreieva@khpi.edu.ua

Fedir Abramov
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
Abramov@khpi.edu.ua

***Abstract* — The article describes in detail the testing and operation principle of an autonomous measuring complex based on the stm32 microcontroller, designed to study the temperature dependence of the electrical resistance of thermistors and determine their parameters: 1) the coefficients included in the Steinhart - Hart equation; 2) thermal time constant characterizing the thermal inertia of the thermistor; 3) temperature coefficient of resistance.**

Keyword s— thermistor, thermistor calibration, Steinhart-Hart equation, thermistor inertia

I. Description of the Measuring Complex

Thermistors are semiconductor resistors with a high temperature coefficient of resistance. Therefore, they are widely used as thermal converters in thermal control and thermal stabilization systems, as well as fire alarms and microwave radiation power meters [1–4]. When operating a thermistor, it is necessary to take into account three main factors: 1) thermistors have a spread of parameters even within the same batch [5]; 2) possible change in the thermoelectric characteristics of the semiconductor resistor during its operation; 3) have a nonlinear temperature dependence of electrical resistance, which is described by the Steinhart-Hart equation [6]:

$$\frac{1}{T} = A + B\ln R + C\ln^3 R\,; \qquad (1)$$

where T is the absolute temperature of the semiconductor resistor; A, B, C - coefficients depending on the parameters of the thermistor and the range of its operating temperatures; R is the thermistor resistance.

Thus, when using a semiconductor resistor as a thermal converter, it is necessary to pre-calibrate and control its parameters during operation. For these purposes, a budget automated measuring complex (fig. 1) was developed based on the STM32F401 microcontroller (MCU), which has a large computing power and RAM, which allows performing various mathematical calculations. The temperature of the test sample is changed using a Peltier element (TEC), on the upper side of which a copper substrate (CS) is fixed. Due to the high thermal conductivity of copper, the thermistor (R) and digital temperature sensor (T) will be in thermodynamic equilibrium at the time of measurement. The thermistor and temperature sensor are fixed on a copper substrate with heat-insulating clips. The underside of the Peltier element is glued with heat-conducting adhesive to a radiator with forced air cooling. The TEC is controlled by a two-channel relay module (RM) and a voltage to current converter (VCC). The relay module sets the operating mode (heating or cooling of the copper substrate) by changing the polarity of the supply voltage of the Peltier element. The temperature of the thermistor at the moment of measurement is determined both by the ambient temperature and by the value of the current flowing through the Peltier element. The amount of this current is controlled by the output voltage of a 12-bit digital-to-analogue converter (DAC) applied to the VCC input.

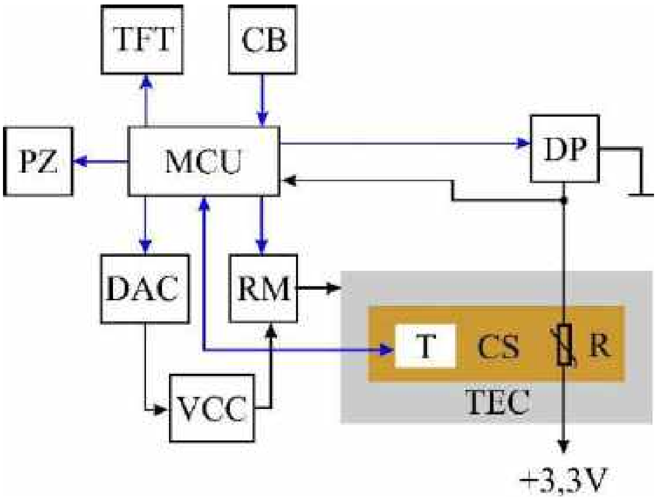


Fig 1. Functional diagram of the measuring complex.

The investigated semiconductor resistor (R) is connected with one output through a digital potentiometer (DP) to a common wire (ground), and with the second, through an inductance to the MCU supply voltage. The voltage from the digital potentiometer through a buffer amplifier with an input resistance of 30 MOhm (not shown in the diagram) is fed to

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

the input of a 12-bit analog-to-digital converter (ADC) of the microcontroller. The thermistor resistance is calculated from the obtained values of voltage and resistance DP. The ADC sampling frequency can be changed in the settings menu in the range from 1 kHz to 1 MHz.

All the necessary information is displayed on a 2.8-inch TFT display with a touchscreen that communicates with the microcontroller via the SPI interface. The unit is controlled by a control unit (CB) consisting of an encoder and two buttons. A passive buzzer (PZ) is used to alert the user with an audible signal. Pairing with a computer is carried out using a USB-UART converter (not shown in the diagram).

The measuring complex can operate in the following modes:

1) *single mode* – thermistor temperature is set, and its resistance is measured;

2) *calibration mode* – automatic measurement of the resistance of the thermistor in the specified temperature range. Approximation of the obtained values by the least squares method and determination of the coefficients A, B, C in equation (1). The display shows experimental points and a graph of the approximation function;

3) *pairing mode* - the measuring complex is controlled and exchanges data with a computer using a specially written program;

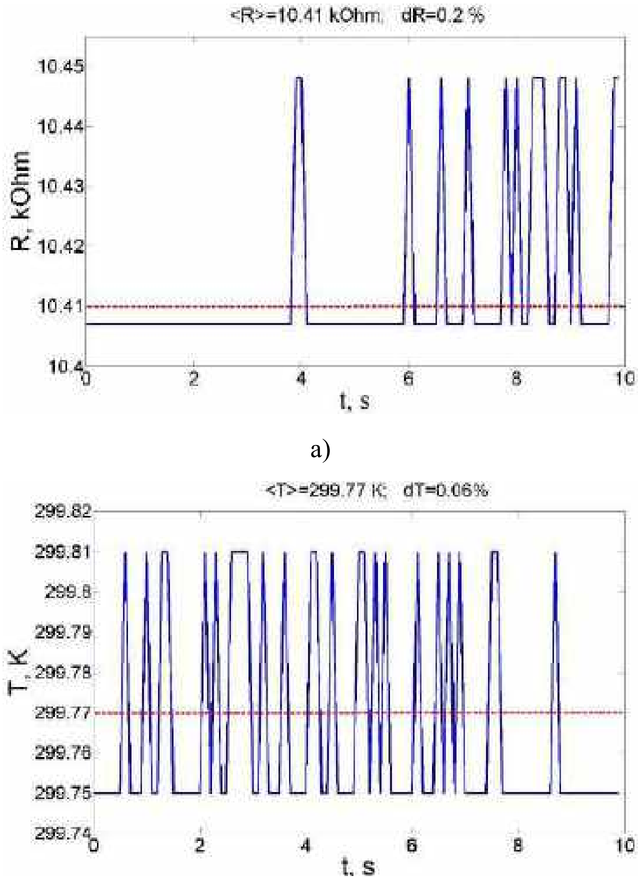
4) *time constant determination mode* - measurement of the thermistor temperature dependence on time when it is freely cooled in air. Least squares linear approximation and determination of the time constant characterizing thermal inertia.

The choice of the operating mode and the setting of the required parameters is carried out using the control unit.

II. Setup Of The Measuring Complex

To determine the parameters of a thermal converter with high accuracy, two conditions must be met: 1) measuring the resistance and temperature of the thermistor with the minimum possible error; 2) carrying out measurements in a stationary state of thermodynamic equilibrium of a digital temperature sensor and a thermistor.

The accuracy of resistance determination depends on the characteristics of the ADC of the microcontroller and on the voltage at the reference resistance connected in series with the thermistor. If the voltage at the ADC input is in the upper half of its scale and the numerical value of the voltage that will be used for calculations is the arithmetic mean of multiple measurements, then the error will be less than 1% (fig. 2a). If the resistances of the thermistor and the reference are very different, then the relative error (dR) increases sharply. Therefore, in the measuring complex, a set of series-connected digital potentiometers acts as a reference resistor, the resistance value of which is adjusted depending on the resistance value of the thermistor in the range from 300 Ohm to 611 kOhm in steps of 10 Ohm.



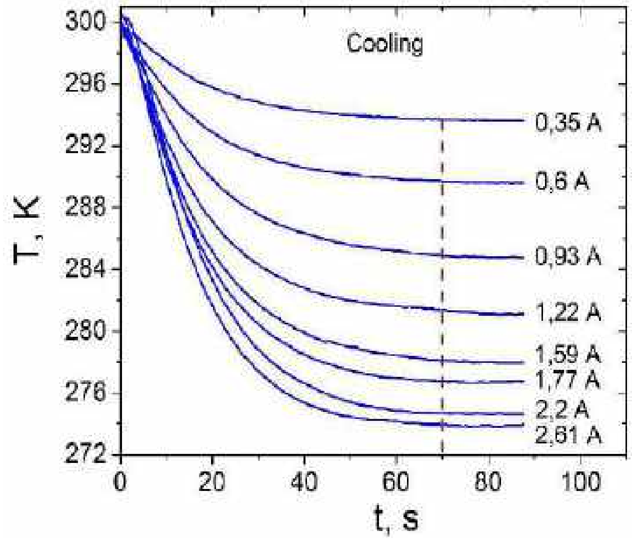
a)

b)

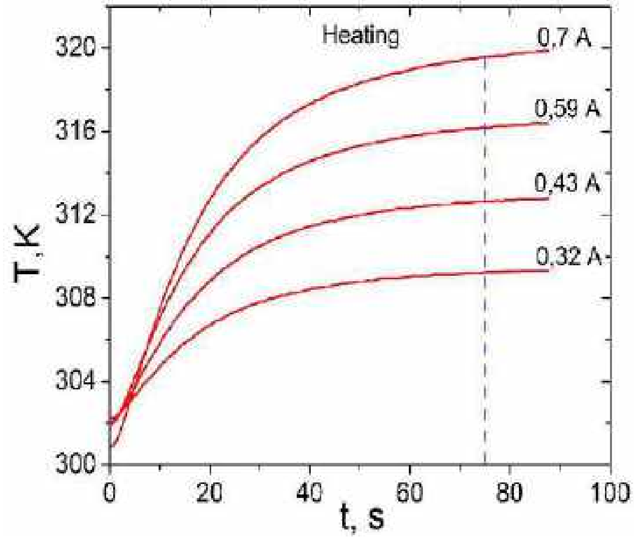
Fig. 2. Time dependences under constant external conditions: a) thermistor resistance; b) digital thermometer temperature.

The measurement error of a digital temperature sensor is determined by its design features. Therefore, the measurement accuracy can only be improved by averaging several temperature measurements (fig. 2b).

The fulfillment of the second condition ensures that the digital thermometer and the thermistor are in a state of thermodynamic equilibrium, and that the temperature and resistance of the thermistor remain constant during their measurement (steady state). The state of thermodynamic equilibrium is achieved due to the copper substrate and good thermal contact of the digital temperature sensor and the thermistor with the copper substrate. The time after which the thermodynamic system ("digital thermometer - thermistor") reaches a stationary thermodynamic state depends on the mode of operation of the Peltier element (heating or cooling) and on the magnitude of the current flowing through it (fig. 3). From fig. 3 shows that the steady state occurs after 70 s in the cooling mode and after 75 s in the heating mode. It was experimentally established that in stationary thermodynamic equilibrium the temperature fluctuates in the range of ±0.1 K, and the change in the resistance of the thermistor does not exceed 0.15 %. Therefore, the measurement of the thermistor parameters during its calibration was performed 75 s after the Peltier element was turned on. The time duration of the stationary state for all dependences presented in Fig. 3 is less than 10 s. Violation of stationarity is observed at currents of more than 1A and is due to insufficient cooling of the radiator.



a)



b)

Fig. 3. Time dependences of the temperature of the copper substrate at various values of the electric current flowing through the Peltier element: a) cooling; b) heating.

The absolute temperature difference (ΔT) between the environment and the copper substrate in the steady state depends on the amount of current flowing through the Peltier element (fig. 4), reaching a maximum value of ΔT≈27 K when cooled and ΔT≈19 K when heated.

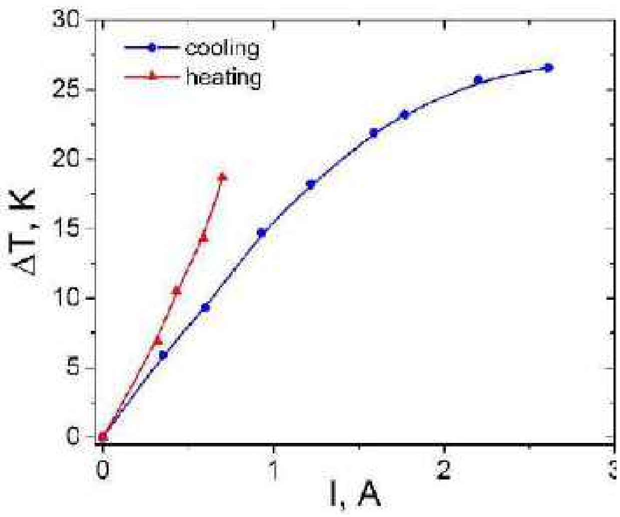


Fig. 4. Dependence of ΔT on the current strength of the Peltier element.

III. Determining Thermistor Parameters

To calculate the coefficients included in the Steinhart-Hart equation, one must select the calibration mode in the settings menu and set the temperature measurement range. This mode determines the resistance of the thermistor in a stationary thermodynamic state at 10 different temperatures. After each individual measurement, the Peltier element is turned off and the next measurement starts after the thermodynamic system "thermistor-digital thermometer" returns to its initial state. Therefore, the total time of the experiment is 25 - 35 minutes. Further, using the least squares method, the microcontroller calculates the calibration coefficients in equation (1) and displays the experimental points and the approximating straight line (Fig. 5), and the user is notified by a buzzer sound signal about the end of the calibration process. The measuring complex allows for automatic calibration of resistance thermistors, which at room temperature range from 350 Ohm to 600 kOhm. In this case, the measurement error of the thermistor resistance will be in the range from 0.15% (350 Ohm) to 2% (600 KOhm).

After calibration, the thermistor can be used as an analog thermometer with high sensitivity and resistance to mechanical stress [7].

In the thermal time constant measurement mode, the thermal inertia of the thermistor is investigated, which is characterized by the time constant τ – this is the period of time during which the temperature difference between the thermistor and the environment ΔT decreases by *e*-times. The time constant depends on the design and size of the thermistor, as well as on the thermal conductivity of the medium in which it is cooled [8].

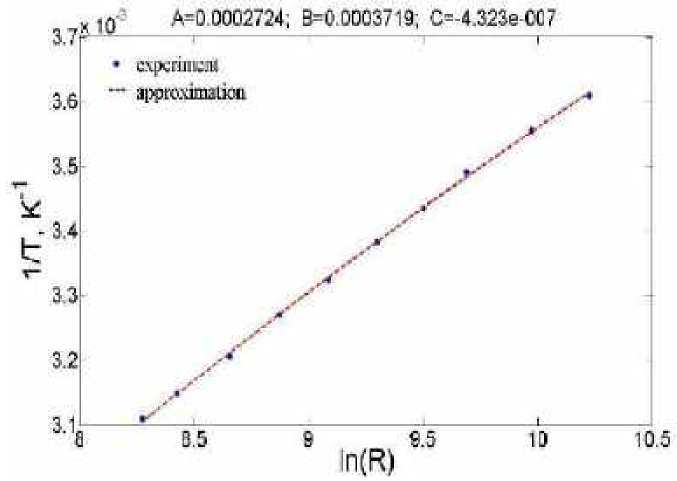


Fig. 5. Graduation of the thermistor.

To determine the parameter τ, you must first enter the values of the coefficients A, B, C using the keyboard, which is shown on the TFT display. Pressing the button on the control unit turns on the Peltier element, which heats the thermistor under test to a temperature of 50°C, and the buzzer notifies the user with an audible signal. Then the Peltier element is turned off, and the thermistor must be moved to a calm air environment. During the cooling process, the resistance of the thermistor is measured (fig. 6) at regular intervals, which are automatically selected depending on the difference in the resistance of the thermistor at room temperature and at a temperature of 50°C.

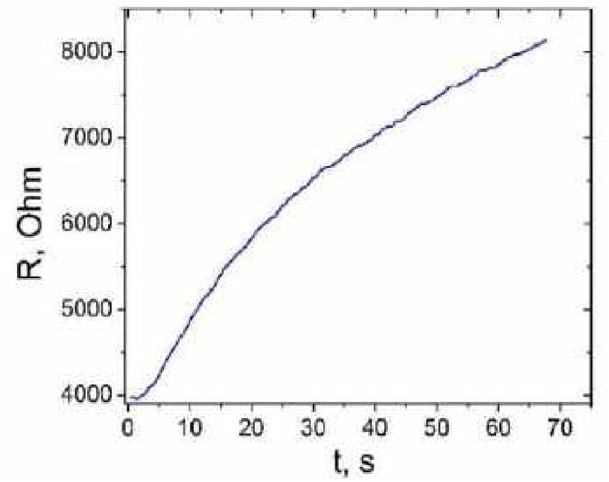


Fig. 6. Time dependence of the thermistor resistance during cooling in a calm air environment.

Using the obtained resistance values, the microcontroller calculates the temperature of the thermistor using equation (1) and performs a linear approximation of the experimental data using the least squares method (fig. 7). The time constant of the thermistor is numerically equal to the cotangent of the angle of inclination of the straight line to the *x*-axis.

In the pairing mode, the temperature coefficient of resistance of the thermistor $\left(\alpha = \frac{1}{R}\frac{dR}{dT}\right)$ is additionally determined in a given temperature range.

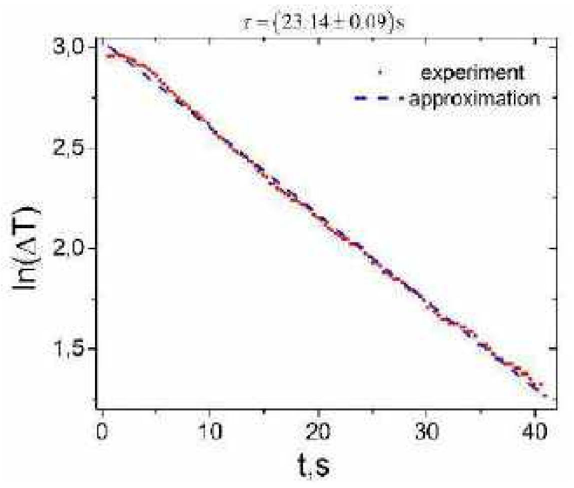


Fig. 7. Determination of the thermistor time constant.

Conclusions

A low-budget (components cost $35) measuring complex has been developed, which allows: to study the temperature dependence of the thermistor electrical resistance; automatically calibrate the thermistor and determine the time constant with high accuracy. The presented complex can be used to develop various devices based on thermistors.

References

[1] A. C. M. Cabrita, R. Mendes, D. A. Quintela, "Thermistor based, low velocity isothermal, air flow sensor," Measurement Science and Technology, 2016, vol. 27, no. 3, p. 035307.

[2] I. Popa, G. N. Popa, C. M. Dinis, A. Iagar, "Temperature-frequency converter made with astable multivibrator and thermistor," Journal of Physics Conference Series, 2021, vol. 1781, p. 012045.

[3] T. Dragos, G. Gherghina, "Evaluation of Thermistors Used for Temperature Measurement in Automotive Industry," Applied Mechanics and Materials, 2018, vol. 880, p. 157.

[4] X. Zhang, C. Wang, J. Ma, G. Ren, "Control and synchronization in nonlinear circuits by using a thermistor," Modern Physics Letters B, 2020, vol. 34, no. 2, p. 2050267.

[5] I. Yang, S. Kim, Y. H. Lee, Y. G. Kim, "Simplified calibration process and uncertainty assessment for sampling large numbers of single-use thermistors for upper-air temperature measurement," Measurement Science and Technology, 2021, vol. 32, no. 4, p. 045002.

[6] J. S. Steinhart., S. R. Hart, "Calibration curves for thermistors," Deep Sea Research and Oceanographic Abstracts, 1968, vol. 15, no. 4, p. 497.

[7] A. D. Harper Smith, D. R. Crabtree, J. L. J. Bilzon, N. P. Walsh, "The validity of wireless iButtons® and thermistors for human skin temperature measurement," Physiological Measurement, 2010, vol. 31, no. 1, p. 95.

[8] M. Tagawa, K. Kato, Y. Ohta, "Response compensation of thermistors: Frequency response and identification of thermal time constant," The Review of scientific instruments, 2003, vol. 74, no. 3, p. 1350 - 1358.

Development and Certification of ECG-pulsometric Device

Tetiana Ryzhenko
Dept. Devices, Technologies and Systems of Contactless Diagnostics Glushkov Institute for Cybernetics of National Academy of Science of Ukraine
Kyiv, Ukraine
tata.ryzhenko@gmail.com

Maksym Mudrenko
Dept. Devices, Technologies and Systems of Contactless Diagnostics Glushkov Institute for Cybernetics of National Academy of Science of Ukraine
Kyiv, Ukraine
mudrenko.m.i@gmail.com

Vitalii Budnyk
Dept. Devices, Technologies and Systems of Contactless Diagnostics Glushkov Institute for Cybernetics of National Academy of Science of Ukraine
Kyiv, Ukraine
vitaliy.budnyk@gmail.com

Mykola Budnyk
Dept. Devices, Technologies and Systems of Contactless Diagnostics Glushkov Institute for Cybernetics of National Academy of Science of Ukraine
Kyiv, Ukraine
budnykmykola@gmail.com

Victor Dehtiaruk
Dept. Data Acquisition Systems Glushkov Institute for Cybernetics of National Academy of Science of Ukraine
Kyiv, Ukraine
degtjaruk@ukr.net

Mykola Dordiienko
G. V. Kurdyumov Institute for Metal Physics of National Academy of Science of Ukraine
Kyiv, Ukraine

***Abstract* — Results of the development and certification process of the device for recording and analysis of electrocardiosignals (ECS) and pulse waves (PW) are presented. The above-mentioned device is designed for recording non-invasive optical signals recorded from the pulse wave caused by the human heart and recording ECS. The hardware includes 2 photometric channels, standard ECG channels, ECG electrodes, electronic (signal processing) unit, laptop and software. The device allows you to examine patients for 10-15 minutes in lying and sitting positions both in clinical conditions and in field conditions. The analysis of registered data is carried out using the PulseWave software, which has passed certification in accordance with international regulations. Hardware and software are components of the portable ECG-pulsometric device CARDIOPULS.**

Keywords* — *computer-aided device, pulsometry, ECG, electromagnetic compatibility, certification.

I. Introduction and Task Statement

Non-communicable diseases are one of the main challenges facing humanity in the 21st century. Cardiovascular diseases (CVD) rank first among non-communicable diseases in the world and are the main cause of death in Western countries and in Ukraine as well. In Ukraine, the mortality from them is 60.2% (according to 2022) of the total mortality [1]. Therefore, the introduction of new modern diagnostic methods is a priority task in solving the problem of early diagnosis of CVD.

The method of simultaneous registration of pulse waves (PW) and ECG combines all the advantages of traditional diagnostic methods with state-of-the-art ones, it makes it possible to detect negative changes in the cardiovascular system in the early stages, which allows solving a wide range of tasks, including preventing diseases or taking measures regarding his warning. Diagnostics allows you to conduct research for a long time. This makes it possible to analyze fairly long sections of recorded recordings and determine the influence of nervous and humoral regulation on the behavior of both the vascular system and the heart [2-6].

One of the integral indicators of the functional state of a person in health and pathological deviations is the dynamics of blood transport in blood vessels, which is reflected in the shape and parameters of PW. With the help of the PW optical sensor, the information about the work of capillaries, elastic vessels and the heart, which is registered from the fingers of the examined person, is processed by the developed software.

Purpose of the work is to create medical device having increased informativeness based on the correlations between ECG diagnostics and pulsometry signals.

II. Structure and Principle of Operation

The portable ECG-photometric complex (Fig. 1) was designed for diagnostics and point-to-point recording in time in certain areas on the surface of the human body:

- the potential difference in ECG leads, which is generated by bioelectric activity of the heart;
- the optical density changes in two PM leads, which is caused by changes in blood volume in the surface layer of biocides, which is generated by biomechanical activity of the heart;
- recording of data in the memory of a PW with further processing of measurement results and visualization in digital and graphic forms.

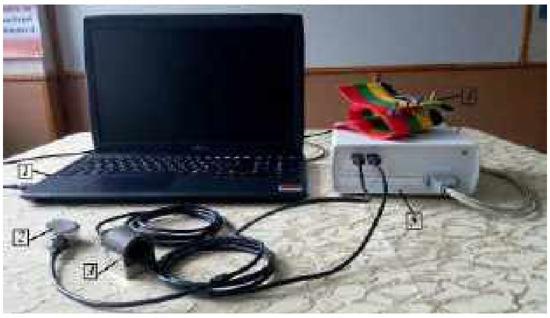


Fig. 1. General view of the complex: 1- laptop; 2, 3 - optical heads; 4 - ECG electrodes; 5 - signal processing unit

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

To obtain full diagnostic information, there is a need for processing and analysis of large data arrays. It suggests the use of modern computer equipment in the composition of photometric instruments, which will allow not only register quantitative values of individual parameters, but also recognize images that allow solving a wide range of problems from the recognition of individual elements of the curves to the diagnosis [6].

Studies are carried out with a non-invasive probing beam of light, without damaging the skin and taking blood.

III. Details of hardware

The complex consists of a signal processing unit, a power supply, a laptop, a software, an ECG cable, 2 optical heads, 4 ECG electrodes. The flow-chart of the device shown in Fig. 2.

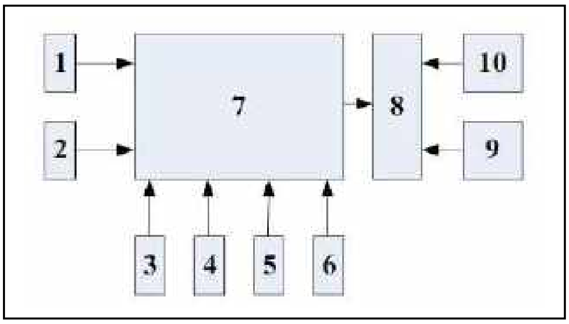


Fig. 2. Flow-chart of the device: 1 – optical finger head; 2 – optical flat head; 3–6 – ECG electrodes; 7 – signal processing unit (SPU); 8 – laptop; 9 – "PulseWave" software;10 – power adapter.

The device includes 2 optical heads (1 – finger and 2 – flat), which are connected via cables to 2 USB inputs Signal Processing Unit (SPU) 7. Signals from the optical heads and ECG electrodes 3–6 enter the SPU for amplifying, filtering and converting signals into digital form. From the SPU, digital signals of eight channels (6 ECG and 2 pulse) enter the laptop 8. The device is connected to a computer via a USB port and works under its control. The data is recorded with the help of the program 9 and written to the PC memory, processed and displayed in the processing program. The laptop is powered by adapter 10, the CPU is powered by the laptop via a USB cable.

The SPU includes the ADAS1000 chip (Analog Devices), which provides amplification, filtering, conversion into a digital 19-bit code and its transmission via a standard SPI serial interface to the AT91SAM512 microcontroller (Atmel). The microcontroller controls the ADAS1000 and transmits the digitized signal to the PC via the built-in USB port. Ultra-bright LEDs are used in the device. The light probe power of the order of 0.1 mW is safe for the human body. Only the direct impact of the light probe on the retina should be avoided. Digital processing of the results allows you to calculate the parameters and assess the state of the vessels, make diagnostic and prognostic conclusions, including the presence and degree of endothelial dysfunction. The operation of the complex is discussed in detail in [8].

The pulsometric method of recording pulse waves combines all the advantages of traditional methods. In addition to high informativeness, it allows conducting research for a long time without affecting the course of theresearched processes. This makes it possible to analyze rather long segments of pulse recordings, which in turn allows to control and give a digital or quantitative assessment of individual components of pulse curves, which are independent in nature and by analogy with the rhythmological approach to the dynamics of cardiac activity [7, 8].

The device has the following operational characteristics:

- the complex is a renewable product;
- according to the potential risk of use, the complex belongs to class IIb, as a non-invasive medical product according to DSTU 4388;
- the complex is a means of measuring technology and is subject to verification once every 1 year.

Finalization of the device according to the requirements of the technical regulation (TR).

All medical equipment used in practice is subject to mandatory registration for certification on the territory of Ukraine. From July 1 2015 technical regulations in the field of medical devices entered into force, one of which is the Technical Regulation (TR) on medical devices dated October 2, 2013 No. 753 [9]. One of the main requirements of this TR is the resistance of the medical product to electrostatic discharges (ESD) and electromagnetic compatibility (EMC). In addition, medical equipment should not create excessive interference.

The requirements of TR for medical products to the ECG-pulsometric complex are taken into account by making changes to the signal processing unit (SPU). To ensure electromagnetic compatibility (EMC) and resistance to electrostatic discharge (ESD), practical solutions recommended in the Intel High-Speed USB Platform Design Guide [5] were applied. Since the first electronic developments, all the parts are subjected to electrostatic discharge. ESD events have peak voltages up to 30kV and therewith they are very dangerous for all kind of integrated circuits.

That there are real concerns regarding the robustness against EMI and ESD is written in Intel's "High Speed USB Platform Design Guidelines" [10]. Intel recommends the usage of a common mode choke for EMI suppressions and another component for protection against ESD pulses.

Two typical schematics for optimized protection of one or two USB ports are shown below: With one TVS diode array WE-TVS you can fully protect two USB lines. All four signal lines as well as the common power supply are well protected (Fig. 3, Fig.4).

The SPU housing was shielded to protect against ESR with a high-conductivity metal shell. EKP-102 grade E organic silicon sealant was used to protect the internal parts and body of the BOS. Ferrite filters are also installed on the SPU cable to ensure filtering of high-frequency noises.

The ECG-pulsometric complex is tested for EMC and resistance to ESD of various magnitudes (+/-2 kV, +/-4 kV, +/-6 kV, +/-8 kV). The tests were carried out in the laboratories of the SE "Ukrmetrteststandard".

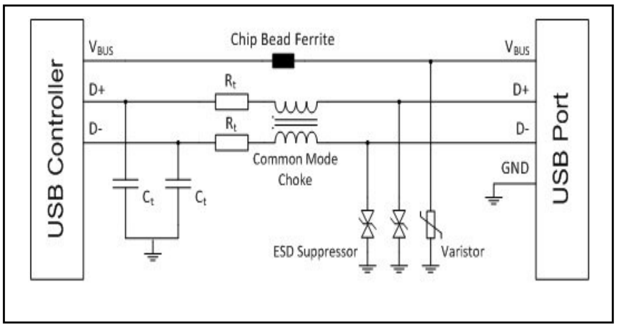


Fig. 3 In contrast to the shielding of the dataline no low capacitive ESD suppressor is necessary at the power supply

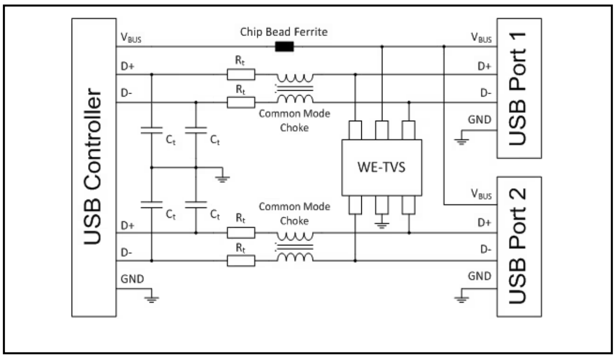


Fig. 4 Two-Port-USB-Interface with ESD-Protection

The results. After carrying out preliminary tests on EMC, it was established that the DC/DC converter, which is part of the BOS circuit, is a source of interference that exceeds permissible parameters. This converter was replaced with a Recom Power RV-0505S, which passed the EMC test [11, 12].

When checking for resistance to ESD effects, malfunctions of the device were observed. In order to increase the resistance of the ECG-pulsometric complex to electrostatic discharge, the body of the ECG connector was replaced with a body with a higher protection class (IP67). Changes have also been made to the software [13].

IV. Software and Certification

The PulseWave software is designed to work with the CARDIOPULS portable ECG-pulsometric complex and provides data reading, storage in the database (DB), data export from the database, data analysis and saving of results in the form of a report when examining patients [14-19].

In addition, the software can work with other digital computer ECG or heart rate monitors that provide the following sets of input signals: 1 PC+4 ECG channels; 2 PC+ 6 ECG channels; 2 PC+12 ECG channels.

Software development planning is performed in accordance with "IEC 62304-2006" Medical equipment software - Software life cycle processes" for security level A (Fig. 5, Fig. 6).

The software implements the following functions:

1) General functions:

• storing the patient's electronic card in the computer memory;

• maintenance of a computer database, which includes signals and a table of indicators, and in which queries are implemented to find the necessary information;
• creation of an electronic report containing ECG curves, a table of ECG indicators and a conclusion.

2) Diagnostics of heart pulse:

• registration of signals
• recording and visualization of the pulse wave;
• automatic determination of the main characteristic points of the pulse wave;
• calculation of diagnostic parameters reflecting the hemodynamic features of the circulatory system, in particular complications of blood vessels;
• automatic determination of base points of the PW channel;
• calculation of the amplitude-time parameters of PW;
• calculation of pulse wave speed.

3) Standard ECG includes: recording of ECG signals; automatic analysis of the amplitude-time parameters of the ECG and heart rhythm disorders; standard HRV analysis; medical conclusion based on the Hannover algorithm. Comprehensive diagnostics of the cardiovascular system - determination of the state of the cardiovascular system based on a combination of ECG and pulsometry.

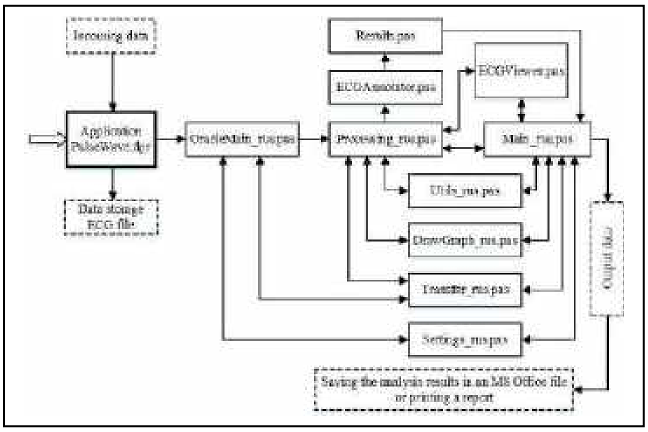
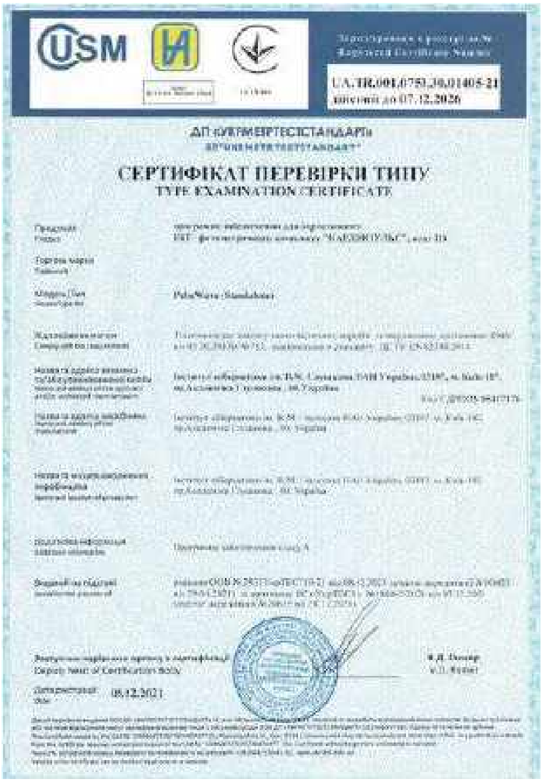


Fig. 5. General structure of the PulseWave software.

The software development plan contains the following information according to IEC 62304-2014 [20]:

- project overview – contains a description of the purpose, scope and tasks of the project. It also defines productsworks provided for by the project;
- project organization – describes the organizational structure of the project team. Thus, everyonethe team member knows his duties and requirements for the work he performs;
- management process – explains the estimated costs and schedule, defines the main stages of the project and describes how the project will be monitored.
- Also, risk management plan were developed including: collection, systematization and analysis of post-production information; identifying the dangers of using the software; risk analysis; risk determination for identified hazards; assessment of risk acceptability; determination of measures aimed at reducing risks; implementation of risk management measures; final risk assessment; validation of risk management tools after implementation; determinationof risks that arise during the implementation

of risk management measures; checking the completeness of the risk assessment; full assessment of residual risk; preparation and approval of the report; preparation of a new version of the plan; evaluation of the risk management process.



UA.TR.001.0753.20.01405-21
дійсний до 07.12.2026

ДП «УКРМЕТРТЕСТСТАНДАРТ»

СЕРТИФІКАТ ПЕРЕВІРКИ ТИПУ
TYPE EXAMINATION CERTIFICATE

Предмет / Product: [illegible]

Торгова марка / Trademark:

Модель (Тип) / Model/Type: PulseWave Standalone

[illegible]

Дата реєстрації / Date: 08.12.2021

Fig. 6. Certificate for PulseWave software.

Conclusions

The improved design of the BOS complex in accordance with the requirements of the technical regulations, taking into account the requirements of international standards regarding tests for electromagnetic compatibility and resistance to electrostatic discharges. EMC and ESR resistance tests were conducted at Ukrmetrteststandard. Based on the results of the tests, changes were made to the design documentation of the ECG-pulsometric complex.

PulseWave software has passed the certification for compliance with the requirements of "IEC 62304-2006" Medical equipment software - Software life cycle processes" in SE "Ukrmetrteststandard".

Acknowledgments

Authors give thanks to National Academy of Science of Ukraine for support within the framework of project VFK 220.30 "Creation of informational and methodological foundations of the technology of combined ECG and pulse diagnostics for its implementation in practical medicine" (2018-2019, manager M.M. Budnyk); project VK 220.31 "Development, experimental operation and introduction into production of biomedical information and diagnostic systems and intelligent sensor devices" (2018-2022, manager M.M. Budnyk); project VF 220.29 "Develop methods and algorithms for processing signals of various nature for the construction of diagnostic and information systems" (2018-2023, manager M.A. Primin).

References

[1] https://opendatabot.ua/open/death-statistics

[2] V. Zharinova, “Endothelial dysfunction as a multidisciplinary problem”, Circulation and haemostasis, vo1.1-2, pp. 9-15, 2015.

[3] J.F. Reckelhoff, “Gender differences in the regulation of blood pressure”, Hypertension, vol.37, pp.1199-1208, 2001.

[4] I. Zaporozhko, V. Zubchuk, E. Nastenko, E. Nosovets, “Age and gender peculiarities of a pulse measurement”, Biomedical Engineering, no.2, pp.48-53, 2011.

[5] A. Zabirnyk, A.Kultaev, “Apparatus pulse diagnostics: theory and practice”, Kharkiv: Noveslovo, 116 p., 2008.

[6] I. Chaikovsky, V. Batushkin, W. Visnevsky, “ECG Universal score system: new instrument for electrocardiogram analysis”, ICCIIDT’16, London, UK, pp. 254-268, 2016

[7] I. Voitovych, V. Korsunskyi, “Intellectual sensors”, Kyiv: Glushkov Institute of Cybernetics NAS of Ukraine, 2007.

[8] V. Dehtiaruk, M. Budnyk, M. Khodakovskyi, M. Mudrenko, V. Mieshkov, “Development of photometric devices for pulsometry”, Proceedings of the VernadskyTauride Natl Univ, Series: Technical Sciences, 2018, no. 5, vol. 29 (68), pp. 39-43.

[9] On the approval of the Technical Regulation on medical devices. Resolution of the Cabinet of Ministers of Ukraine dated October 2, 2013, №753 - URL: https://zakon2.rada.gov.ua/laws/show/753-2013-%D0%BF.

[10] The Protection of USB 2.0 Applications - URL: https://www.we-online.com/components/media/o109030v410%20AppNotes_ANP002_TheProtectionOfUSB20Applications_EN.pdf

[11] V. Degtiaruk, M. Khodakovsky, M. Budnyk, V. Budnyk, M. Mudrenko, Ya.. Tymoshenko. Development of metrological maintenance of photometric devices for pulsometry. Метрологія і прилади. - №4. -2019. – pp. 10-16. DOI:10.33955/2307 2180(4)2019.1016

[12] Mudrenko M.,Dehtiaruk, V., Budnyk, V., Ryzhenko T., Budnyk, M. Finalization and testing of the ECG-photometric complex according to the technological regulations. 2 International scientific and practical conference «Information systems and technologies in medicine» (ISM–2019) November 28–29, 2019 Kharkiv, Ukraine: Collection of scientific articles – pp.200-202.

[13] Budnyk, V., Ryzhenko T., Mudrenko M.,Budnyk, M. Finalization of the software of the ECG-photometric complex for compliance assessment. Ibid. – pp. 182-184.

[14] Ryzhenko T., Dehtiaruk, V., Budnyk, V., Budnyk, M., Chaikovsky, I., Sofienko, S.Development and Studying Value of Method of Non-Invasive Pulsometry(2019) 2019 IEEE 39th International Conference on Electronics and Nanotechnology, ELNANO 2019 - Proceedings, art. no. 8783839, pp. 512-517. DOI: 10.1109/ELNANO.2019.8783839

[15] K-M. Delavar, I. Zaporozhko, V. Zubchuk, O. Skoryk, V. Tkachenko, “Dynamic pulse diagnostics”, Electronics and communications, vol.2, Kyiv: AVERS, pp.252-257, 2009.

[16] Ryzhenko T., Budnyk, V., Budnyk, M., Development and evaluation of compliance of the software of the ECG-photometric complex. Cybernetics and computer technologies. 2021. №3. pp. 115-128. https://doi.org/ 10.34229/2707-451x.21.3.10

[17] M. Budnyk, V. Dehtiaruk, Method of registration of pulse waves, Utility model patent UA 140747, A61 B5/0295, G01N 21/25, public 10.03.2020, Bulletin "Industrial Property" № 5, p. 4.20.

[18] V. Dehtiaruk, M. Budnyk, M. Khodakovskyi, M. Budnyk, V. Ryzhenko, T., Mudrenko, I. Chaikovsky, Development of an ECG-photometric complex in accordance with the requirements of the technical regulations for medical devices. Proceedings of the VernadskyTauride Natl Univ, Series: Technical Sciences, 2019, no. 4, vol. 30 (69), Part 1, pp. 29-33.

[19] Budnyk, M., Dehtiaruk, V., Budnyk, V., Ryzhenko T.,ChaikovskyI. Technology of combined ECG and pulsometric diagnostics. Database "Technologies of Ukraine"/ RCT technology registration card, state register№0620U000119, 23.11.20. UkrINTEIhttp://rkt.ukrintei.ua/rkt/rkt_ 26505 e0494662534f633586941b77d0c.pdf

[20] DSTU EN 62304:2014. Medical device software. Software life cycle processes provision (EN 62304:2006, IDT). https://www.iso.org/obp/ui/#iso:std:iec:62304:ed-1:v1:en (accessed 09/16/2021).

Development of 9-channel Magnetocardiograph CARDIOMOX MCG-9

Vitalii Budnyk
Dept. of sensor devices, systems and technologies of contactless diagnostics
Glushkov Institute of Cybernetics of NAS of Ukraine
Kyiv, Ukraine
vitaliy.budnyk@gmail.com

Mykola Budnyk
Dept. of sensor devices, systems and technologies of contactless diagnostics
Glushkov Institute of Cybernetics of NAS of Ukraine
Kyiv, Ukraine
budnykmykola@gmail.com

Volodymyr Sosnytskyy
Dept. of sensor devices, systems and technologies of contactless diagnostics
Glushkov Institute of Cybernetics of NAS of Ukraine
Kyiv, Ukraine
sosnytskyy1@gmail.com

Xie Feng
Suzhou Cardiomox Ltd
Suzhou, Jiangsu, P.R. China
xie.feng@cardiomox.cn

Tao Miao
Suzhou Cardiomox Ltd
Suzhou, Jiangsu, P.R. China
tao.miao@cardiomox.cn

Shouzhang Zhu
Suzhou Cardiomox Ltd
Suzhou, Jiangsu, P.R. China
shouzhang.zhu@cardiomox.cn

Pavlo Sutkovyi
Dept. of sensor devices, systems and technologies of contactless diagnostics
Glushkov Institute of Cybernetics of NAS of Ukraine
Kyiv, Ukraine
pavsutk@meta.ua

Yurii Minov
Dept. of sensor devices, systems and technologies of contactless diagnostics
Glushkov Institute of Cybernetics of NAS of Ukraine
Kyiv, Ukraine
yu54minov@gmail.com

Pavlo Spylovyy
Dept. of sensor devices, systems and technologies of contactless diagnostics
Glushkov Institute of Cybernetics of NAS of Ukraine
Kyiv, Ukraine
shpylovy@gmail.com

Maksym Mudrenko
Dept. of sensor devices, systems and technologies of contactless diagnostics
Glushkov Institute of Cybernetics of NAS of Ukraine
Kyiv, Ukraine
mudrenko.m.i@gmail.com

Yevhenii Melnyk
Dept. of sensor devices, systems and technologies of contactless diagnostics
Glushkov Institute of Cybernetics of NAS of Ukraine
Kyiv, Ukraine
evgm1991@gmail.com

Wenming Ji
Oxford Cardiomox Ltd
Oxford, United Kingdom
wenming.ji@cardiomox.com

***Abstract* — Description of novel generation of magnetocardiography device Magnetocardiograph CARDIOMOX MCG9 has been presented. Device is intended for measurements of magnetic fields induced by the human heart for medical diagnose of cardiology diseases. Hardware includes 9-channel SQUID-magnetometer, single-channel reference electrocardiograph, control and processing unit, and automatic electromechanical system for patient scanning. Device can perform examination of the patients within 15 minutes at unshielded clinical location. Novel design and advanced methods for device adjusting, testing and certification were proposed.**

Keywords — electronic medical device, magnetocardiograph, computer-aided data acquisition system, adjusting, calibration, sensors

I. Introduction

The Magnetocardiography (MCG) is well-known non-invasive tool in a world for register and analysis of super-weak magnetic field signals in the area of the human heart. It can be used in research institutions, cardiologic clinics, and other medical hospitals [1]. It can help to identify the presence of cardiovascular changes or anomalies in the heart earlier than other diagnostic methods. So, based on results of observation, a medical doctor can propose in the form of a recommendation a set of certain medical exercises or other propositions, for example, additional analyzes, to help the patient feel better which can cause to diseases in the future and improve the patient's health condition [2].

The such systems are very sensitive, because they use Superconducting quantum interference detectors (SQUIDs) to register super-weak magnetic signals. It is necessary very low temperature for these sensors, so they are placed in a cryostat filled with liquid helium [3].

It is known that to receive more accurate data during the patient's study systems should be more complex with more quantity of measuring channels.

During 2002-2004 year (project 2187, Science and Technology Center in Ukraine (STCU) [4], funded by EU and Canada), the 4-channel MCG system CARDIOMAG have been made. The device includes 8 channels: 4 MCG channels to register heart currents, 3 reference ones measuring the magnetic noise, and ECG channel. During 2009-2011 year, next generation of 4-channel device CARDIOMAGSCANER was developed and installed at National military medical clinical center «Main Military Clinical Hospital» (Kyiv) within the Project STCU #4719 [5,6].

The examination of patients with the help of the developed MCG system is carried out by the user or doctor who was pre-prepared and trained to use the system, studied the Instruction for user and Instructions for the operation of the device, passed the appropriate theoretical and practical exam and received the appropriate certificate from the developer or the owner of device.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

Essence of work is development of MCG device in order to provide reliable working in clinic locations without any magnetic or radiofrequency shielding room. Another aspect is introducing worldwide new technical regulations according to which procedures for assessment of conformity should be implemented for medical devices. In this connection, demands on safety, risk management, electromagnetic compatibility, and usability are essentially increases. That is why quality of devices must be essentially improved. Detail technical requirements for the 9-channel device have been developed in 2011 [7].

Task of the work is development of MCG device which will have 9 MCG channels, automatic bed to move the patient, improved software, and, of course, device should stable operate in magnetic unshielded environment.

II. Overall Description of Device

The “Magnetocardiograph “CARDIOMOX MCG9” device is intended for use as tool which non-invasively measures and displays the magnetic signals produced by the electric currents in the heart. It allows reliable revealing the early stages of ischemic heart disease due to the detection of ischemic changes in viable myocardium, to evaluate the quality of treatment with drugs used for patients, etc. [8]. Diagnostics is carried out by Clinician using the graphical results and quantitative MCG characteristics.

The measuring device consists of the main components presented at following flowchart at Fig.1:

1. Automatized patient scanning system (APSS) including gantry and bed;
2. Control and processing electronics;
3. Multi-channel recorder of cardiomagnetic signals based on SQUID-sensors;
4. Workstation comprise of PC or laptop, monitor, printer;
5. Software package for device control and adjustment, data acquisition and processing.

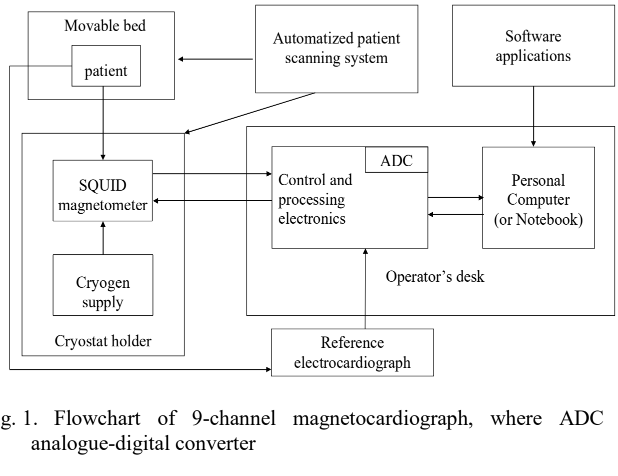

Movable bed
patient
Automatized patient scanning system
Software applications
SQUID magnetometer
ADC
Control and processing electronics
Personal Computer (or Notebook)
Cryogen supply
Operator’s desk
Cryostat holder
Reference electrocardiograph

Fig. 1. Flowchart of 9-channel magnetocardiograph, where ADC – analogue-digital converter

The device (for general view, see Fig. 2 and Fig.3) registers with no direct contact the MCG of the heart processed by means of the electronics. Patient’s ECG is being recorded simultaneously in one of the standard leads and is used as a reference signal. From the ECG output, the signal is also transmitted to the electronics. MCG&ECG analog signals are transmitted to the ADC converting those into the digital form. Control of the magnetometer is made either from the PC or from the electronics (off-line mode).

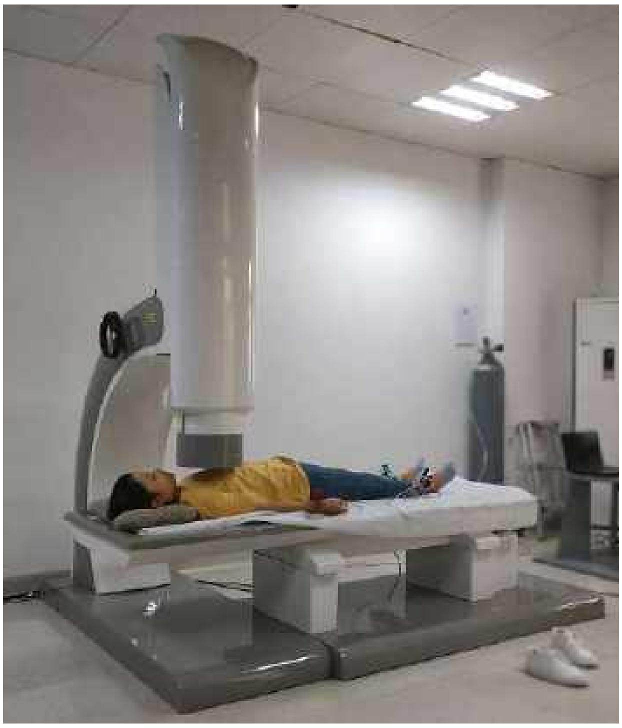


Fig. 2. General view of the device (gantry and bed)

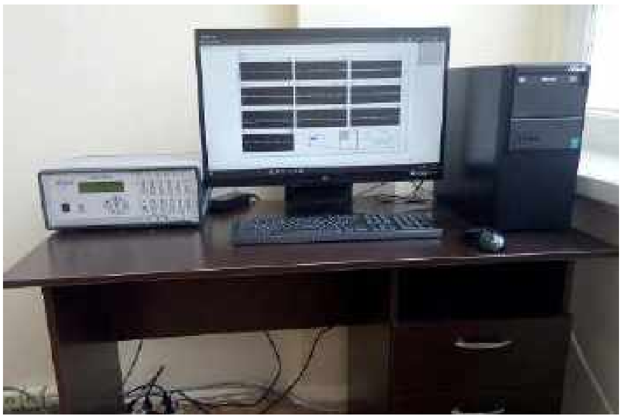


Fig. 3. General view of the device (operator workplace)

The Magnetocardiograph has operational attributes shown in Table I.

TABLE I. DEVICE OPERATIONAL ATTRIBUTES

No.	Name of attribute	Value
1.	Productivity (patients per hour)	4
2.	Operational life time	≥ 5 year
3.	Shelf (storage) life time	≤ 5 year
4.	Warranty service life time	1 year after putting into operation
5.	Technical service period	1 year
6.	Helium life cycle after refilling	5 days if working day ≤ 8 hours

No.	Name of attribute	Value
7.	Mode of operation	continuous
8.	Trouble-free time	≥ 2 000 hours if working week ≤ 5 days×8 hours
9.	Electric power consumption	≤ 600 VA

III. Hardware

The main parts of device hardware are: Control and processing electronics (ECP, see Fig.4), Multi-channel Recorder of Cardiomagnetic Signals, Cryogenic supply, Automatized patient scanning system (APSS), Operator workstation (PC or laptop, monitor, and printer).

ECP includes unit for power supply, control and data processing (CPU) with ADC, reference ECG, 3 Multiplexers for Recording Channels (MRCs) (see Fig. 5), and cables set (see Fig. 4). Cables set includes 7 cables among which are cables for connection of MRCs, ECG, service, control signals to APSS, connection USB to PC, and 2 power cables. The ECP is controlled and regulated by user from PC. Requirement for ECG device is simple mode without any signal processing. ADC is designed to convert signals from analogous form into digital one and input it into PC.

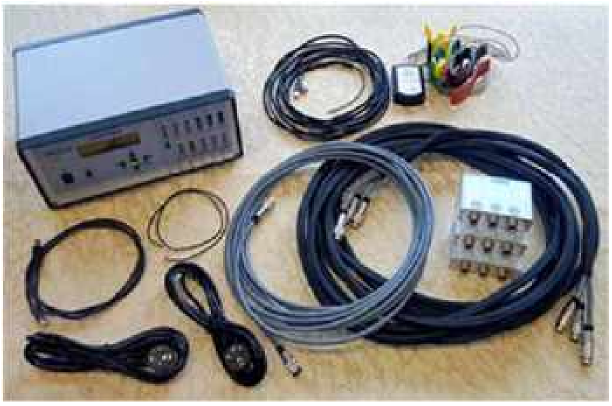


Fig. 4. General view of Control and processing electronics

The operation of ECP is follows. Signals from 9 MCG channels are transmitted to 3 MRCs. MRC group signals and transmit those via the MCG data cable to the CPU. The CPU processes MCG signals and transmits them to the PC.

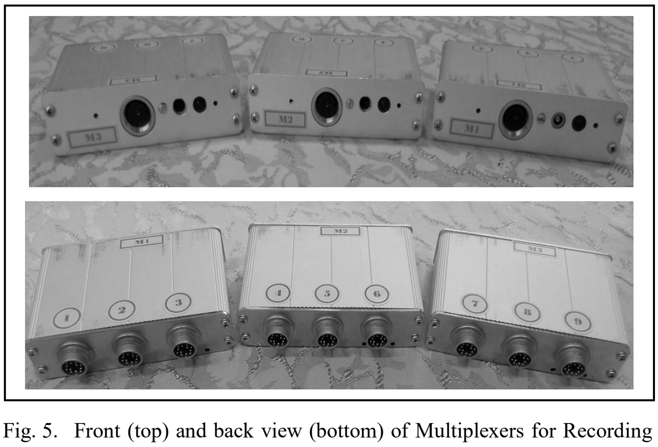


Fig. 5. Front (top) and back view (bottom) of Multiplexers for Recording Channels with 9 connectors to MCG probes

Control and processing unit (CPU). The CPU transmits measuring signals from individual measuring probes to the computer. Flowchart for Control and processing unit is shown in Fig. 6 and front view – in Fig. 7. Control signals from computer reach the CPU and might be passed to the sensors. For every channel a bar graph is available (see 5-13, Fig. 7), showing the signal level. A bar graph of ECG channel signal level is also available (see item 14, Fig. 7).

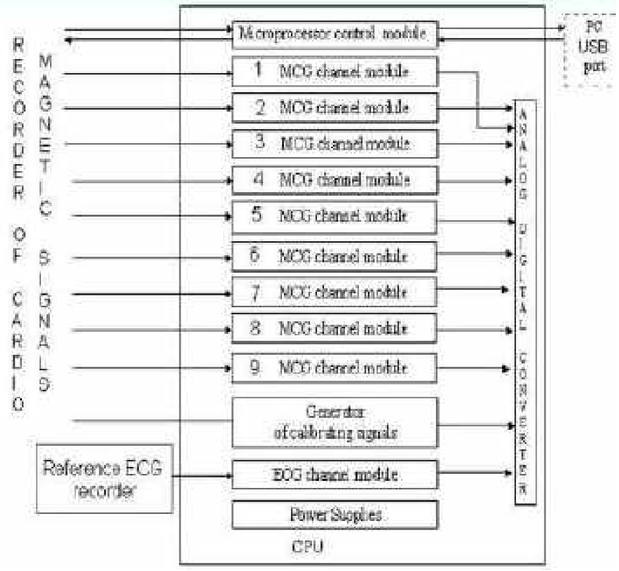


Fig. 6. Flowchart of Control and processing unit

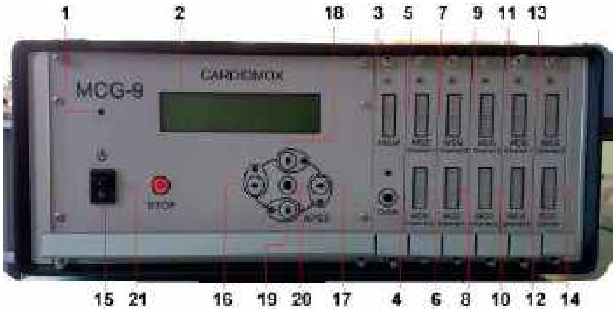


Fig. 7. Front view of Control and processing unit

CPU has digital output of measuring signals to the computer via USB cable. The CPU is switched via the system's main switch. As you can seefrom Fig. 5, CPU also have: 1 – Power lamp showing if the device is switched on or off; 2 – LCD display; 3 – bar graph of helium level; 4 – button for activation helium measurement; 15 – CPU power switch; 16-20 – movement APSS to the left, right, forward, backward, and fixation of position; 21 – APSS emergency stop button.

Multi-channel Recorder of Cardiomagnetic Signals (MRCS). Recorder is intended for simultaneous registration of the vertical component of the heart magnetic field at 9 spatial points located at the 3x3 cm grid with 4 cm step. Current SQUID-sensor "CS blue" (Supracon, Jena, Germany) is mounted on a cryogen probe made of non-magnetic materials with low thermal expansion coefficient. MRCS consists of 9 signal probes set fixed in and installed in thermostat. The helium level sensor is also embedded to the thermostat; flowchart is shown at Fig. 8. Photo of general view of MRCS is presented at Fig. 9.

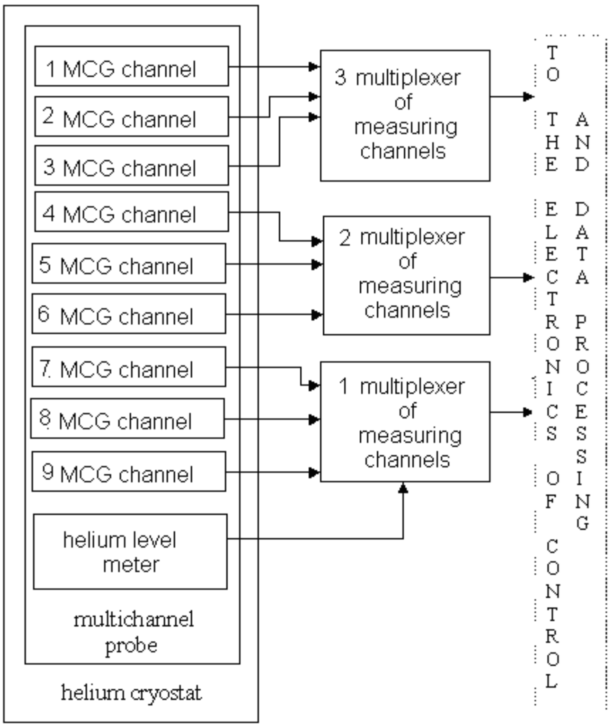


Fig. 8. Flowchart of 9-channel Recorder of Cardiomagnetic Signals

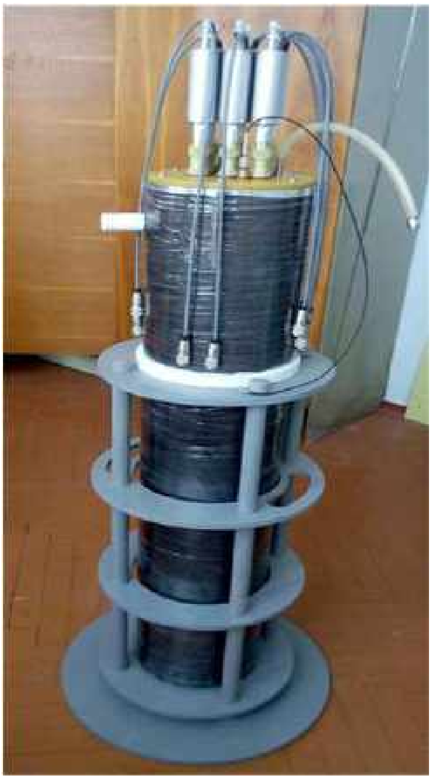


Fig. 9. View of 9-channel Recorder of Cardiomagnetic Signals

Cryogenic supply. It includes helium thermostat and refilling device. Flexible refiller was made that allows fill helium into thermostat without removing thermostat from gantry. Thermostat is made from non-magnetic fiberglass passing magnetic field with the smallest coefficient of helium atoms diffusion for vacuum inside the thermostat not to degenerate. Fiberglass magnetotransparent thermostat (FMT) LH-13.1-B (Cryoton, Troitsk, Russia), have the following parameters: length – 1100mm, external diameter – 227 mm, internal diameter – 155 mm, helium capacity – 13,1 liters, weight – 14 kg, helium evaporation rate < 2 liters / day.

Automatized patient scanning system (APSS) is intended for displacement of the patient under the 9 antennae to cover all 36 spatial points and for fixing the thermostat at a certain distance from the heart. APSS consists of bed and gantry. The patient bed (see Fig. 10, without the top part and plastic details) is located below the measuring sensors and serves as a bed during the measurement.

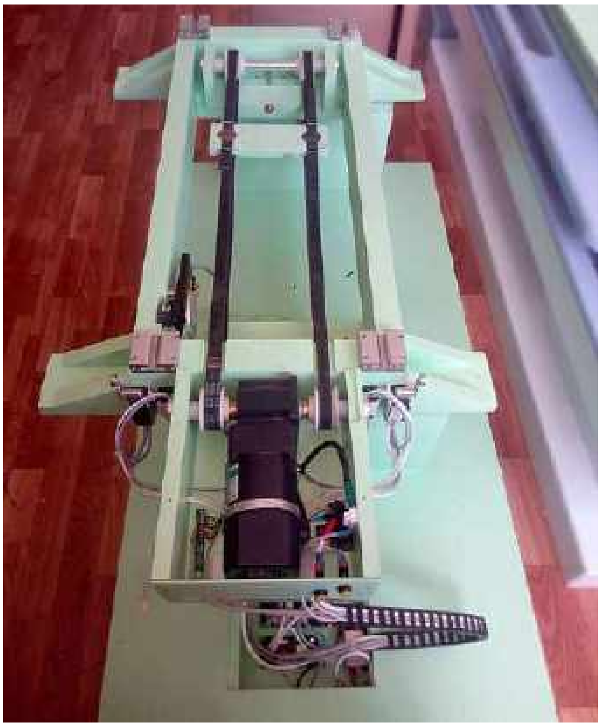


Fig. 10. Design of moving patient's bed (without top platform)

The bed made by plywood and covered by plastic, on which there is a washable mattress. For a comfortable position of the patient a head part is included in the delivery, so that the head of the patient can be lying comfortably. The bed can be moved horizontally, so that the patient can easily get on and off the bed outside the measuring sensors. When the patient is lying relaxed on the patient bed and the ECG electrodes have been attached to his hands and legs. As default, bed is moved automatically and PC-controlled. For the manual operation the buttons on the CPU shall be used.

The starting position of the patient is important for the reproducibility of the measurement, because this way it is guaranteed that the measurement always starts from the same position. Subsequently 4 measuring points are adjusted in a 2x2 matrix and at each measuring point the magnetic field is recorded for approximately 30-60 seconds. When the measurement has been completed the bed is moved back into the initial position in order to facilitate the patient's getting off the bed.

The gantry (see Fig. 11) is located above the bed and holds measuring capsule with the sensors. The sensors during the measurement are placed above the patient's thorax. In order to guarantee a good quality of the measuring signals, the distance between the patient's thorax and thermostat must not exceed few cm. For that purpose the measuring capsule can be shifted up and down manually with a moving wheel. The gantry is also made by plywood and covered by plastic details for better general view.

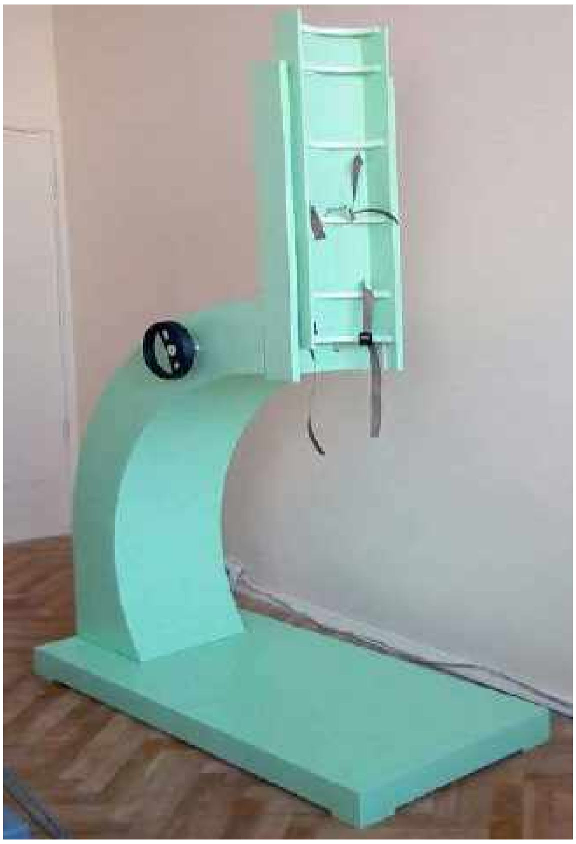


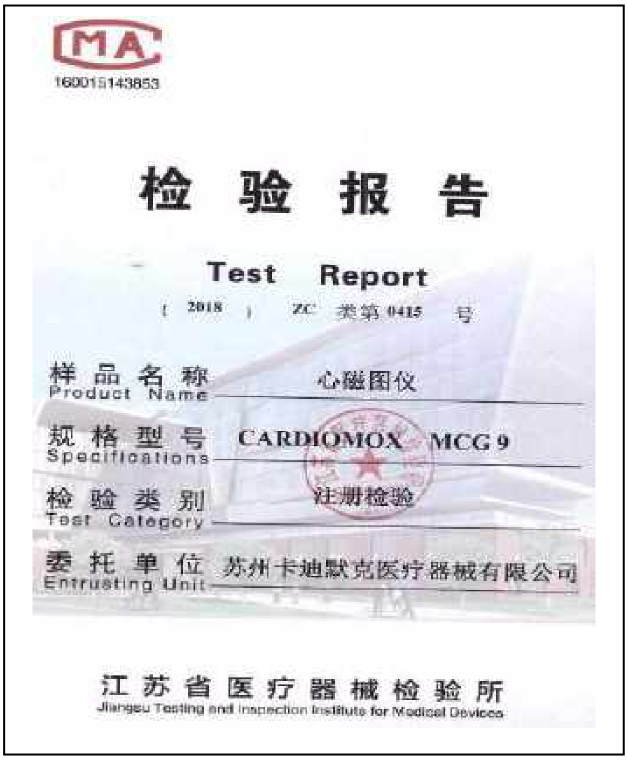
Fig. 11. General view of gantry

Advanced design of MCG channels [9] and frame of input antennas (2nd order gradiometers) based on carbon-filled composite plastic [10] have been developed in order to increase reliability against mechanical vibrations and stability of initial imbalance due to thermo-cycling (Yu.Minov, M.Budnyk, V.Liakhno, O.Shopen, O.Kivirenko "Thermostable superconductive magnetic gradiometer", China Patent Granted CN107430174B, 2020).

IV. Certification

Certification of MCG device prototype was carried out in China in a few stages. Firstly, at 2017-2018 year technical tests of the system were carried out and test report received. The tests were carried out by Jiangsu Testing and Inspection Institute for Medical Devices in 2018 (see Test Report at Fig.12). In result, we validate main technical quality criteria that are shown in the Table II. Main technical parameters of MCG channels were collected in Table III.

At the other stages Certification of MCG device prototype was continued in next year. As a result, Conformity Report (see Fig. 13) and Conformity Assessment Certificate (see Fig. 14) were obtained.



CMA
160015143853

检 验 报 告

Test Report

(2018) ZC 类第 0415 号

样品名称 Product Name: 心磁图仪

规格型号 Specifications: CARDIOMOX MCG 9

检验类别 Test Category: 注册检验

委托单位 Entrusting Unit: 苏州卡迪默克医疗器械有限公司

江苏省医疗器械检验所
Jiangsu Testing and Inspection Institute for Medical Devices

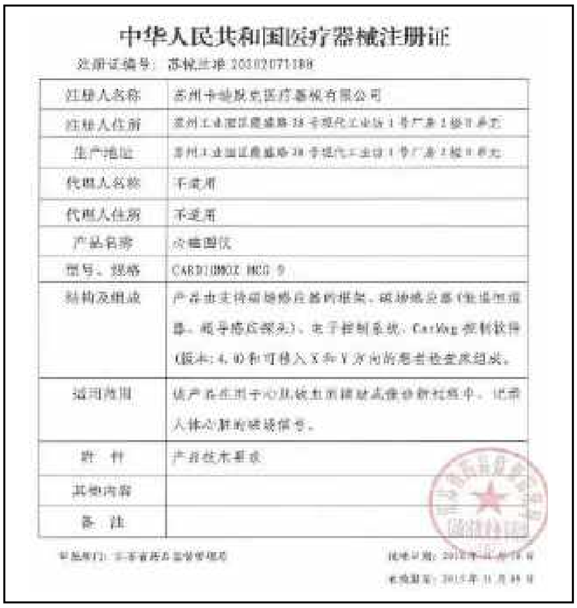
Fig. 12. First page of the Test Report for device CARDIOMOX MCG9

TABLE II. Essential technical quality criteria

No.	Name of criteria	Value
1.	Size of scanning area	200 × 200 mm
2.	Accuracy of patient positioning	2 mm
3.	Dynamic range	83,1 dB
4.	Transfer factor (voltage/magnetic field)	3,9 mV/pT
5.	Magnetic field resolution	0,36 pT

TABLE III. Main parameters of MCG channels

No.	Name of parameter	Value
1.	SQUID operating mode	standard DC
2.	Magnetic flux resolution (white noise, i.e. per 1 Hz above 1/f cut-off frequency)	$\leq$ 3 mk$\Phi_0/\sqrt{\text{Hz}}$ $\Phi_0=2\text{x}10^{-15}$ Wb – magnetic flux quantum
3.	Slew rate	$\geq 2\text{x}10^5$ Φ_0/sec
4.	Frequency band at level –3 dB	20 kHz
5.	Output noise voltage	$\leq$ 2 mV
6.	Antenna type	axial wire-wound 2nd order gradiometer
7.	Registered field component	vertical
8.	Spatial layout of channels	9 nodes of 3x3 grid with 8 cm step



中华人民共和国医疗器械注册证

注册证编号：苏械注准 20202071188

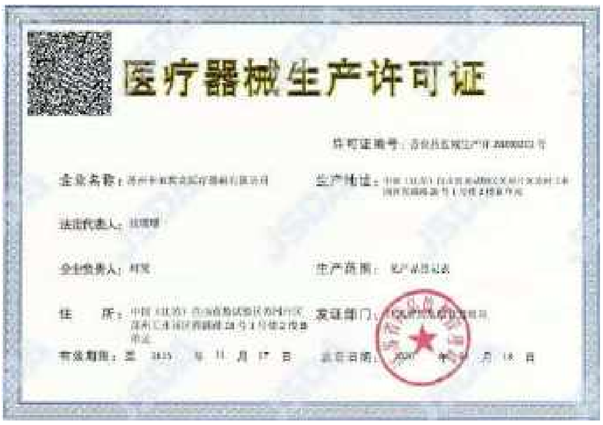
注册人名称	苏州卡迪奥克医疗器械有限公司
注册人住所	[illegible]
生产地址	[illegible]
代理人名称	不适用
代理人住所	不适用
产品名称	心磁图仪
型号、规格	CARDIOMOX MCG 9
结构及组成	产品由支持磁场感应器的框架、磁场感应器（包括杜瓦器、超导感应探头）、电子控制系统、CarMag 控制软件（版本：4.0）和可移入 X 和 Y 方向的患者检查床组成。
适用范围	[illegible]
附　件	产品技术要求
其他内容	
备　注	

审批部门：江苏省药品监督管理局

批准日期：[illegible]

有效期至：2025年11月09日

Fig. 13. Conformity Report for device CARDIOMOX MCG9



医疗器械生产许可证

许可证编号：[illegible]

企业名称：[illegible]

生产地址：[illegible]

法定代表人：[illegible]

企业负责人：[illegible]

生产范围：[illegible]

住　　所：[illegible]

发证部门：[illegible]

有效期限：至 2025 年 11 月 17 日

发证日期：2020 年 [illegible] 月 18 日

Fig. 14. Conformity Assessment Certificate

CONCLUSIONS

Five MCG devices have been developed at V.M. Glushkov Institute of Cybernetics of NAS of Ukraine within the P624 STCU project. They were installed in China medical institutions (4 devices) and UK (1 device). Devices are working now and personnel setup work around each device to ensure a continuous flow of people or patients who need such examination and the possibility of supplying a certain volume of liquid helium in a folded schedule.

Some improvements of design of 2nd order gradiometers [9, 10] and calibration procedure have ensured stable and reliable operation of device at unshielded location in presence of large urban magnetic disturbances. Devices were assembled, adjusted, and calibrated. The device consists of the 5 following main components: Automatized patient scanning system including gantry and bed; Electronics for control and processing; Multi-channel recorder of cardiomagnetic signals based on SQUID-sensors; Workstation comprise of PC or laptop, monitor, printer; Software package for device control and adjustment, data acquisition and processing.

Advantages of developed device are novel design, computer-aided moveable automated bed, 9 MCG channels, and absence of reference vector magnetometer and noise compensation. The device, filled with liquid helium, can work during no more than 7 working days if every day device is used no more than 8 working hours.

Devices have been certified by Jiangsu Testing and Inspection Institute for Medical Devices in 2018. Also, the MCG devices were tested in medical clinics in Beijing and Harbin, and after device passed clinical testing in 2019. As a result, it was obtained a Certificate of conformity assessment as medical device in 2020.

ACKNOWLEDGMENTS

This device has been developed within the framework of STCU (Sci&Tech Center in Ukraine) project P624 Development of method and apparatus for 3D imaging of electrophysiological activity of biological tissues P624(2014-2016), P624a (2016-2017), P624b (2017-2018), P624c (2018-2023), partner company Oxford Cardiomox Ltd, UK. Authors from Glushkov Institute of Cybernetics give thanks for financial support to Suzhou Cardiomox Ltd (Suzhou, Jiangsu, P.R. China) and personally to Company Director Xie Feng.

REFERENCES

[1] H. Koch, “Recent advances in magnetocardiography“. Journal of Electrocardiology, 37, 2004. pp. 117-122 https://doi.org/10.1016/j.jelectrocard.2004.08.035.

[2] B.A. Steinberg, A. Roguin, S.P. Watkins III, et al. “Magnetocardiogram recordings in a nonshielded environment – reproducibility and ischemia detection”. Ann. Noninvasive Electrocardiol., vol. 10. 2005. – pp. 152-160. https://doi.org/10.1111/j.1542-474x.2005.05611.x.

[3] R. Fagaly. “Superocnducting quantum interference device instruments and applications”. Rev. Sci. Instrum. Vol. 77 (011101). 2006. p. 1-45. https://doi.org/10.1063/1.2354545.

[4] I. Voytovych, M. Budnyk, Y. Minov, L. Artemenko, E. Paslavsky, P.Shpilyovy, “First Ukrainian multi-channel magnetocardiograph: Assembling and testing”, Proceedings of the 2nd IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2003, art. no. 1249507, 2003, pp. 16-19. DOI: 10.1109/IDAACS.2003.1249507.

[5] I. Chaykovskyy, M. Najafian, M. Budnyk, S. Martynenko, A.Dovbysh, O. Kovalenko, “Development of pattern recognition method for diagnosis of myocardial ischemia and noncoronarogenic myocardial diseases based on current density distribution maps”, In: 17th International Conference on Biomagnetism: Advances in Biomagnetism – Biomag2010. IFMBE Proceedings, Supek, S., Sušac, A. (eds), vol 28. Springer, Berlin, Heidelberg. 2010, pp. 424-427. https://doi.org/10.1007/978-3-642-12197-5_101.

[6] M. Budnyk, O. Zakorcheny, V. Koshelnyk, V. Budnyk, V. Lukashyk, Y. Minov, P. Sutkovyi, T. Ryzhenko, “Improvement of small-channel MCG system for unshielded environment”. Ibid. - pp. 66-69. DOI: 10.1007/978-3-642-12197-5_11.

[7] M. Budnyk, V. Sosnytskyi, I. Voitovych, V. Maiko, Yu. Minov, P.Sutkovyi, O. Zakorchenyi, T. Ryzhenko, V. Budnyk, “Development of 4-channel cardiomagnetic scaner and technical requirements for 9-channel scaner to diagnose the heart abnormalities”. In Proc. of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS'2011, 1, art. no. 6072717, 2011, pp. 91-96. DOI: 10.1109/IDAACS.2011.6072717.

[8] W. Ji. Oxford Cardiomox – a New Generation of Magnetocardiography (NG-MCG). https://innovation.ox.ac.uk/wp-content/uploads/2014/08/Oxford-Cardiomox-Wenming-Ji.pdf.

[9] M. Budnyk, Y. Minov, V. Lyakhno, V. Desnenko, A. Linnik, O. Shopen, “Development of improved superconductive axial gradiometers for biomagnetic SQUID applications” Low Temperature Physics, 44 (3), 2018, pp. 233-237. DOI: 10.1063/1.5024543.

[10] M. Mudrenko and Y. Melnyk, “Creation and testing of modernized measuring probes with superconducting antennas”, Computer tools, nets and systems, No. 17, pp. 20-25 (2018) (in Ukrainian).

Software development for Magnetocardiograph CARDIOMOX MCG-9

Vitalii Budnyk
Dept. Devices, Technologies and Systems of Contactless Diagnostics
Glushkov Institute for Cybernetics of NAS of Ukraine
Kyiv, Ukraine
vitaliy.budnyk@gmail.com

Mykola Budnyk
Dept. Devices, Technologies and Systems of Contactless Diagnostics
Glushkov Institute for Cybernetics of NAS of Ukraine
Kyiv, Ukraine
budnykmykola@gmail.com

Xie Feng
Suzhou Cardiomox Ltd
Suzhou, Jiangsu, P.R. China
xie.feng@cardiomox.cn

Yurii Frolov
Dept. Devices, Technologies and Systems of Contactless Diagnostics
Glushkov Institute for Cybernetics of NAS of Ukraine
Kyiv, Ukraine
fread@voliacable.com

Pavlo Sutkovyi
Dept. Devices, Technologies and Systems of Contactless Diagnostics
Glushkov Institute for Cybernetics of NAS of Ukraine
Kyiv, Ukraine
pavsutk@meta.ua

Yevhenii Udovychenko
Igor Sikorsky Kyiv Polytechnic Institute,
Kyiv, Ukraine
yevhenii.udovychenko@gmail.com

***Abstract* — Detail description of software for novel medical device, i.e. Magnetocardiograph CARDIOMOX MCG-9has been presented. Above device is intended for noninvasive measurements of magnetic signals induced by the human heart. Software package called CARMAG, Version 4.1, can be divided onto 2 parts based on its function. The 1st part is intended to input of signals, storage of measurement results in the database and digital preprocessing of signals and includes modules Package Shell, Input9, Embedded software, BDE, and Database. The 2nd one is intended to selection of a cardiocycle interval for examination, inverse problem solution, calculation of quantitative indexes, report generating, and includes modules Topology, Currents, Matlab, and PDF Creator. Software can perform examination of the patients within 15 minutes. Novel methods for data processing were utilized. Device is aimed to diagnose of heart and vessel diseases.**

Keywords — computer-aided device, medical device, software, magnetocardiogram, control, data acquisition

I. Introduction and Task Statement

The global burden of non-communicable diseases is one of the main challenges facing humanity in the 21st century. It has undermined socio-economic development and very negative impact on the demographic situation.

Cardiovascular diseases (CVD) rank first among non-infectious diseases. CVD are the leading cause of death in Western countries and Ukraine too. In Ukraine mortality from them is 66.5% of total mortality. About 8 million Ukrainians suffer from coronary heart disease, each year they register about 50 thousand myocardial infarctions. The number of strokes in Ukraine is 13 times higher than in Europe [1]. Therefore, the introduction of new modern diagnostic methods is a priority task in solving the problem of early diagnosis of CVD. Modern medicine prefers non-invasive methods, one of which is magnetocardiography (MCG) [2].

The MCG method records super-weak magnetic fields and has few advantages compare to traditional diagnostic methods. MCG diagnostics provides an opportunity to detect changes in the cardiovascular system at the early stages, which allows you to solve a wide range of problems, including preventing the disease or taking measures to prevent it. In addition to high informativeness, diagnostics allows you to carry out research over a long period of time without affecting the patient. Some approaches to analyze MCG allow you to reconstruct current distribution into the human heart and is promising for diagnostics [3-4].

Early, authors were a part of team that developed a 4-channel computer-aided MCG device with manual-driven patient positioning system [5-7]. The purpose of the work was to create an advanced software for 9-channel device having automatized patient bed and enabling stable working at real urban hospitals in noisy environment.

II. Device Structure

The “Magnetocardiograph “CARDIOMOX MCG9” device is intended for use as tool which non-invasively measures and displays the magnetic signals produced by the electric currents in the heart. Diagnostics is carried out by Clinician using the graphical results and quantitative MCG characteristics.

The measuring device consists of the main components presented at following flowchart at Fig.1:

1. Automatized patient scanning system (APSS) including gantry and bed;
2. Control and processing electronics;
3. Multi-channel recorder of cardiomagnetic signals based on SQUID-sensors;
4. Workstation comprise of PC or laptop, monitor, printer;
5. Software package CARMAG for device control and adjustment, data acquisition and processing.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

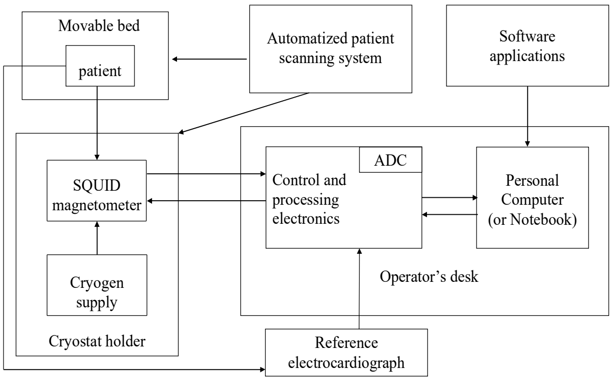

Movable bed
patient
Automatized patient scanning system
Software applications
SQUID magnetometer
ADC
Control and processing electronics
Personal Computer (or Notebook)
Cryogen supply
Operator's desk
Cryostat holder
Reference electrocardiograph

Fig. 1. Flowchart of 9-channel magnetocardiograph, where ADC – analogue-digital converter

The device (see Fig. 1) registers the magnetic field from the human heart without direct contact processed by means of the electronics. Patient's ECG signal is being recorded simultaneously in one of the standard leads and is used as a reference signal. From the ECG output, the signal is also transmitted to the electronics. MCG&ECG analog signals are transmitted to the ADC converting those into the digital form. Control of the magnetometer is made either from the PC or from the electronics (off-line mode).

MCG signal lies in super-weak picoTesla range and can be measure only by high-sensitive recorder. Superconducting quantum interference detectors (SQUIDs) to register super-weak magnetic signals are used. It is necessary very low temperature for these sensors, so they are placed in a cryostat filled with liquid helium.

Main part of device is Multi-channel Recorder of Cardiomagnetic Signals, which is intended for simultaneous registration of the vertical component of the heart magnetic field at 9 spatial points located at the 3x3 cm grid with 4 cm step. Current SQUID-sensor "CS blue" (Supracon, Jena, Germany) is mounted on a cryogen probe made of non-magnetic materials with low thermal expansion coefficient. 9 signal probes installed into cryostat with liquid helium.

Such high-tech device can reliable work in clinic locations without any magnetic or radiofrequency shielding room. Detailed description of device and structure of its main parts can be found at [8].

III. General Software Structure

Software structure diagram, showed at Fig.2, describe the relationship between composition modules, modules functions and modules relationship, and the relationship between modules and external interface. According to software diagram it can be divided onto 2 parts taking into account its function:

1) Input of signals, storage of measurement results in the database and digital preprocessing of signals (modules Package Shell, Input9, Embedded software, Borland Database Engine (BDE), Databases).

2) Identification of a cardiocycle interval for examination, inverse problem solution, calculation of quantitative indexes and report generating (modules Topology, Currents, Matlab Runtime Compiler, PDF Creator).

The first part intends to hardware control, data acquisition, distinguish MCG cardiocycles and signal de-noising. It is intended for use in scanning mode. This preliminary processing is carried out once after each examination data have been inputted.

The 2nd part, the visualization of spatial distributions of magnetic field and current sources and temporal dynamic of changes is carried out. It is intended for use in processing mode. This secondary processing is carried out off-line after preprocessing of each group of patients.

Third party software and needed hardware environment are collected into Table I.

PC Workstation controls and adjusts medical device hardware through the software, installed on personal computer. PC workstation and medical device connecting each other via USB cable between PC and Control and Processing Unit (CPU).

IV. General Software Structure

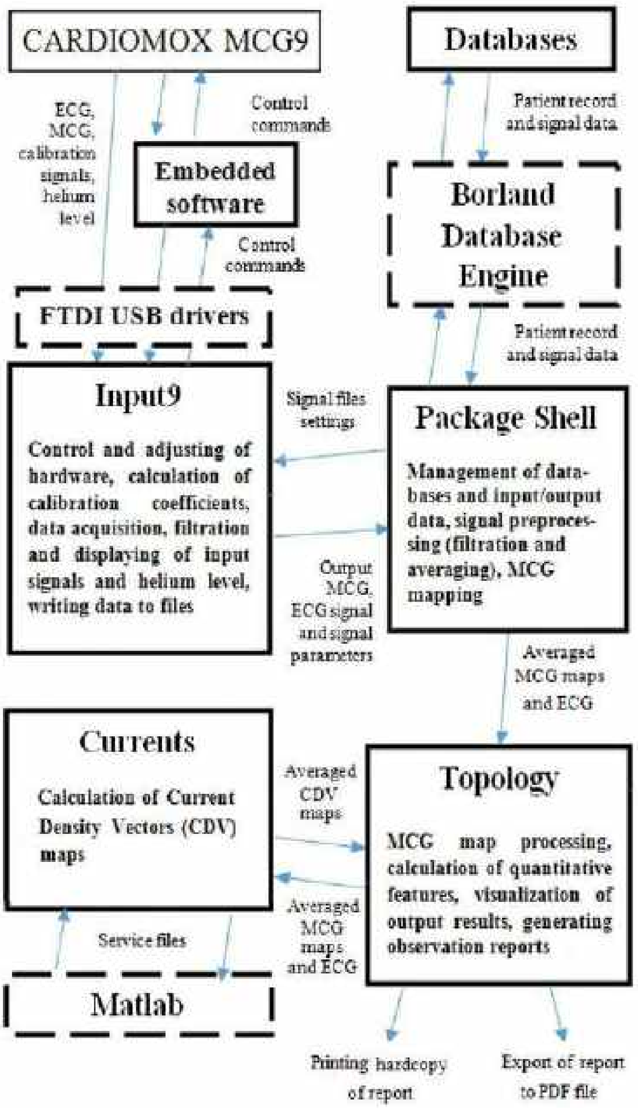

CARDIOMOX MCG9
Databases
ECG, MCG, calibration signals, helium level
Control commands
Patient record and signal data
Embedded software
Borland Database Engine
Control commands
FTDI USB drivers
Patient record and signal data
Input9
Control and adjusting of hardware, calculation of calibration coefficients, data acquisition, filtration and displaying of input signals and helium level, writing data to files
Signal files settings
Package Shell
Management of data-bases and input/output data, signal preproces-sing (filtration and averaging), MCG mapping
Output MCG, ECG signal and signal parameters
Averaged MCG maps and ECG
Currents
Calculation of Current Density Vectors (CDV) maps
Averaged CDV maps
Topology
MCG map processing, calculation of quantitative features, visualization of output results, generating observation reports
Averaged MCG maps and ECG
Service files
Matlab
Printing hardcopy of report
Export of report to PDF file

Fig. 2. Software package CARMAG: bold box – firmware, dotted box – system, necessary or supporting items developed by third parties, thick box – hardware components

TABLE I. EXTERNAL INTERFACE

No.	Component	Version	Manufacturer
		Software	
1.	OS Windows	10Pro 64-bit	Microsoft Corporation
2.	FTDI USB drivers	2.12.26	Future Technology Devices Intern. Limited
3.	BDE	5.1.1.1	Borland Software Corporation
4.	Matlab Runtime Compiler	Ver. 9 Update 01	Matworks
5.	PDF Creator	6.7.0	Geek Software GmbH.
		Hardware	
6.	Workstation	PC ASUS BM1AD1	ASUSTeK Computer Incorporated
7.	Monitor	27'' LCD	ViewSonic Corporation
8.	Device	CARDIOMOX MCG9	Suzhou Cardiomox Ltd.
9.	Printer	Color laser	Hewlett-Packard

General view of main window of CARMAG software are presented at Fig. 3. Here we can see the raw data of one MCG measurement loaded from database.

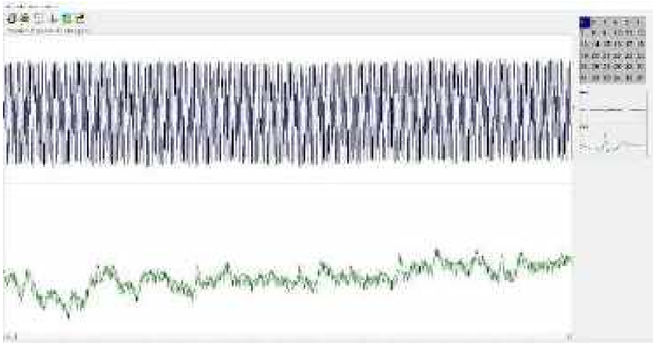


Fig. 3. General view of CARMAG software window

After averaging and filtering we can press tab “Cardiocycle” (see Fig.3) in button menu and, in result, 36 averaged MCG signals are displayed as shown at Fig. 4.

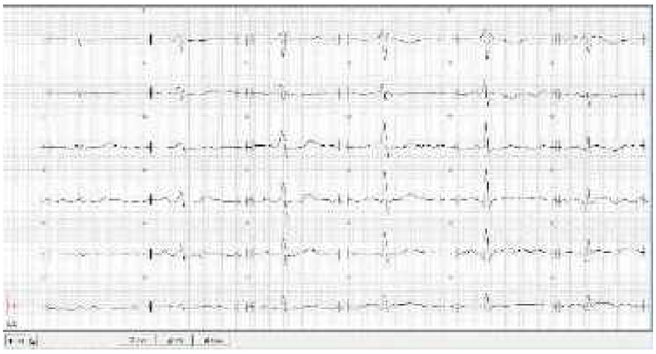


Fig. 4. View of window with 36 averaged MCG signals after filtering

Control and adjust commands go to CPU and then CPU sends them to embedded software of medical device parts (see Fig.5). In the opposite direction data signals from embedded software goes to CPU. After that the data signals through analog-digital converter of CPU are inputted to PC and recorded to the database with help of Package Shell.

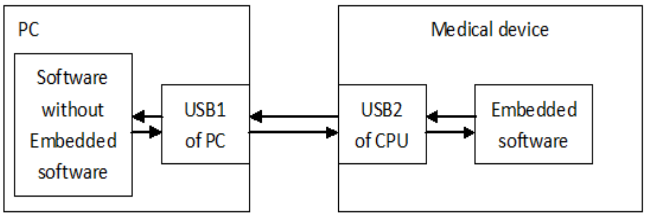


Fig. 5. Physical connection relationships scheme of hardware and software

After the examination data was recorded to database it can be processed and the results can be printed to the paper using Color laser printer or they can be saved as PDF report using PDF Creator program to send it later to patient email.

V. EMBEDDED SOFTWARE

The embedded software consists of seven software modules locating into respective microcontrollers. Data exchange and control commands are carried out through appropriate interfaces. The links between software modules are shown at Fig. 6.

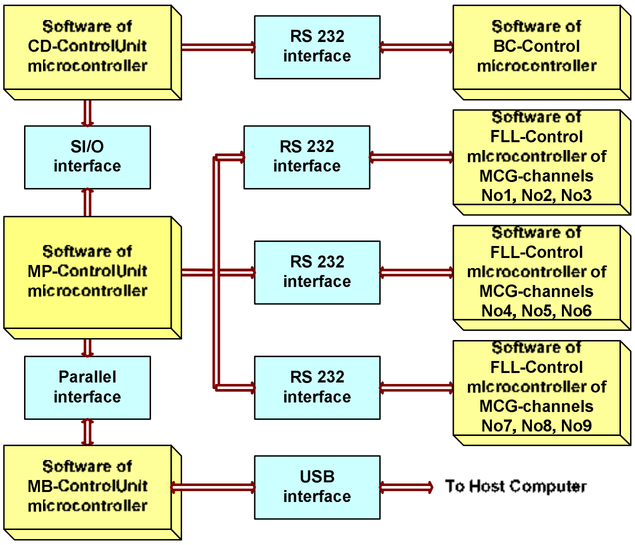

Software of CD-ControlUnit microcontroller
RS 232 interface
Software of BC-Control microcontroller
SI/O interface
RS 232 interface
Software of FLL-Control microcontroller of MCG-channels No1, No2, No3
Software of MP-ControlUnit microcontroller
RS 232 interface
Software of FLL-Control microcontroller of MCG-channels No4, No5, No6
Parallel interface
RS 232 interface
Software of FLL-Control microcontroller of MCG-channels No7, No8, No9
Software of MB-ControlUnit microcontroller
USB interface
To Host Computer

Fig. 6. Structure of the microcontrollers embedded software

The software of microcontroller MP-ControlUnit serves as a central microcontroller. The software generates control signals for the measurement channels, indicates the level of liquid helium in the cryostat, exchanges commands and data with the software of microcontrollers MB-ControlUnit, CD-ControlUnit and FLL-Control.

The software of microcontroller MB-ControlUnit reads data from ten ADCs, transmits them to the host computer, and also participates in the transfer of data between the software of microcontroller MP- ControlUnit and the host computer.

The software of microcontroller CD-ControlUnit communicates with the software of microcontroller MP-ControlUnit, reads the status of the control buttons, displays information on the symbol display and controls the software of BC-ControlUnit microcontroller.

Software of three microcontrollers FLL-Control controls the operation of nine measuring probes (each of which controls three measuring probes). And also the software of first microcontroller FLL-Control reads the level of liquid helium in the cryostat. Via RS-232 interface, the software of microcontroller FLL–Control exchanges commands and data with software of microcontroller MP-ControlUnit.

The software of microcontroller BC-Control is designed to control the movement of the patient's positioning bed. To do this, it generates control signals for electrical drives and reads the position information of the patient's positioning bed. Via RS-232 interface, the software exchanges commands and data with software of microcontroller CD-ControlUnit.

In the device we use tree types of microcontrollers: M30624FGAFP, M30281FAHP, R5F21134FP. Manufacturer – RenesasElectronics Corporation. First MC have next parameters: 16-bit, flash, 16MHz, microcontroller, PQFP100, 14x20 mm, 0.65 mm pitch, plastic, QFP-100. The second MC have next parameters: 16-bit, single-chip microcomputer, M16C FAMILY, Tiny series. The third MC have next parameters: 16-bit, R8C CISC, 16 kb Flash, 3.3V/5V, 32-Pin, LQFP. [9]

VI. Acquisition Module

The software module INPUT9 is intended to register input signals and control of device. Main units of module are presented at Fig. 7.

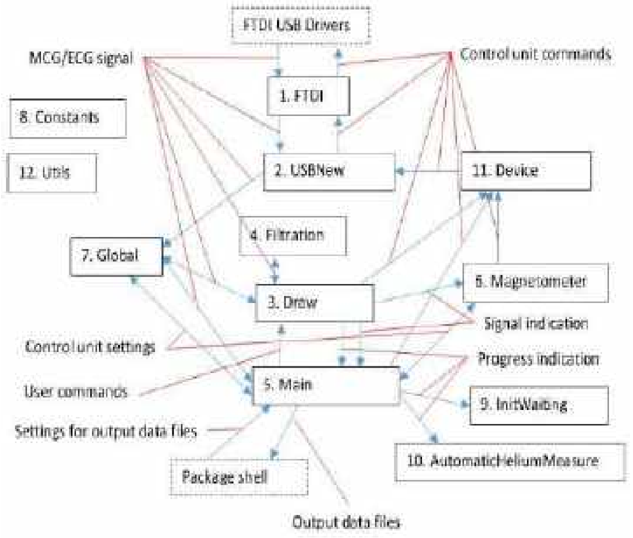


Fig. 7. Design of software module INPUT9

Function of software units:

1. FTDI – interface for a work with FTDI USB Drivers (interface, initialization, adjusting, data reading/ writing, etc).
2. USBNew–real-time thread for buffering and sending of commands to a control unit and receiving of MCG/ECG signal.
3. Draw – real-time thread for real-time signal preprocessing and indication and providing of a calibration, a helium level checking and signal recording.
4. Filtration – unit for signal HFF and LFF filtration (for visualization only).
5. Main – main program window, which allows (for an user) to control an MCG/ECG signal indication/preprocessing/saving, a calibration, a helium level checking, to change control unit settings, to receive settings for output data files, to call item 6.
6. Magnetometer – a program window, which allows (for an user) to control magnetometer settings.
7. Global – unit, which contains/saves/loads program settings, contains an input signal buffer and other global variables (reading/ writing of configuration parameters into appropriate file).
8. Constants – unit, which contains program constants
9. InitWaiting – window, which indicates a progress of control unit initialization/finalization.
10. AutomaticHeliumMeasure –window, which indicates a progress of initial helium level checking.
11. Device – unit, which converts a commands for control unit to device interface format and set they to a buffer.
12. Utils – unit, which contains general service utilities.

External software requirements: FTDI D2XX Drivers (http://www.ftdichip.com/Drivers/D2XX.htm) and CarMag Package Shell.

General view of INPUT9 module window during test examination are presented at Fig. 8.

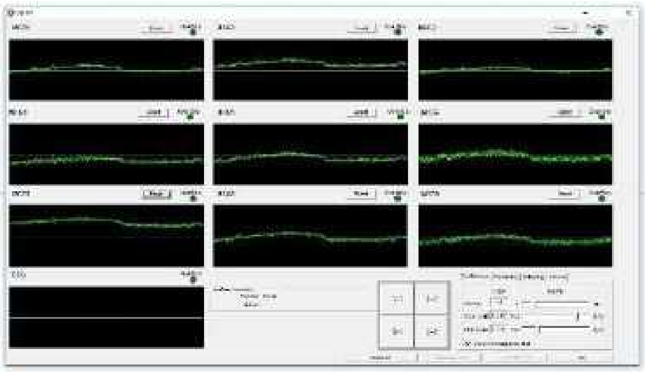


Fig. 8. General view of INPUT9 module window

VII. Package Shell

The main file CarMag.exe has 4 main functions:

- manage of databases;
- input/output data;
- signal preprocessing (ECG analysis, MCG filtration and averaging);
- MCG mapping.

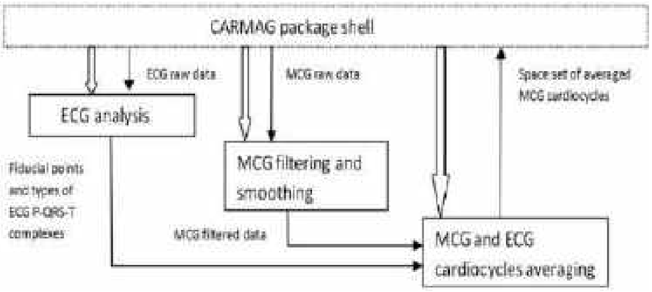

CARMAG package shell
ECG raw data
MCG raw data
Space set of averaged MCG cardiocycles
ECG analysis
MCG filtering and smoothing
Fiducial points and types of ECG P-QRS-T complexes
MCG filtered data
MCG and ECG cardiocycles averaging

Fig. 9. Design of the data preprocessing

Signal preprocessing includes three main stages (Fig.9):

1) ECG rhythm and morphologic analysis. It is necessary for synchronization of MCG data in the different space positions and effective averaging of MCG cardiocycles;

2) MCG filtering and smoothing. It is necessary for suppressing of external noise and improving of the MCG measurements quality;

3) MCG and ECG averaging of normal beats. The set of 36 averaged MCG cardiocycles is the main output data after preprocessing. This information is used for next MCG processing.

Three main parts of preprocessing are realized as three software units “ECG”, “MCG” and “Averaging”. Input data for the preprocessing stage is organized as internal arrays with raw ECG and MCG records for all channels and measurements positions. Output data with averaged cardiocycles is written to the binary file *.b0a, which is used for next processing, i.e. magneticfield mapping.

VIII. Calculation of Magnetic Maps

Next software module “Topology” (Fig.10), which has important functions: MCG map processing, calculation of quantitative features, Visualization of output results and generating of reports.

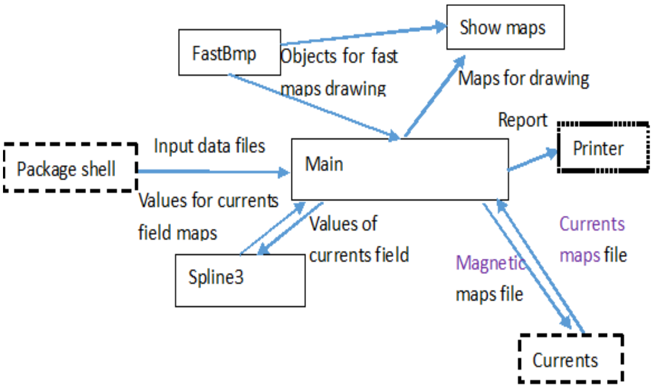


Fig. 10. Design of the module TOPOLOGY

As we can see from Fig.10 the module has next units: Main – main form, which calls external application for getting of current vectors maps, shows and brows combined vector/scalar current vectors maps, calculates, shows and prints report information; ShowMaps – a form, which shows a set of all combined vector/scalar current vectors maps; Spline3 – a unit, which provides a scalar magnetic map interpolation for its displaying; FastBmp – a unit for fast work with 2D images.

Used External libraries: AlgoLib. URL: http://www.alglib.net/. License: GPL 2+. Used unit: Spline3 (sources). It is used for interpolation of 2D-map of magnetic field.

General view of module “Topology” window presented at Fig. 11.

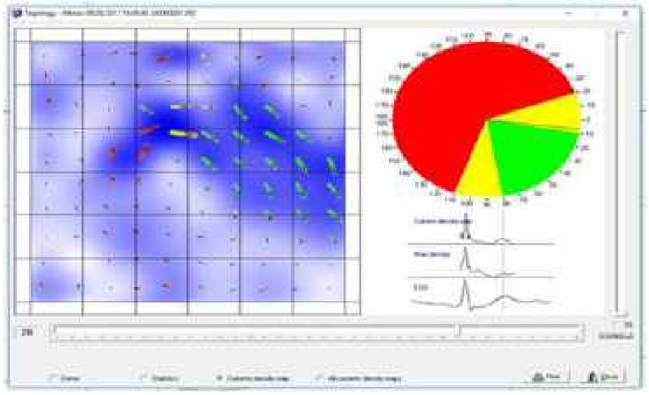


Fig. 11. Vizualisation of current density maps for selected time interval

Used files:

1. General header file. Binary file (*.h*), which contains a basic information of record (date, time, patient name and sex, calibration coefficients).
2. Detailed header file. Text file (info.ti), which contains package shell version, patient name, sex, age and birthday, measurement date/time.
3. Fields.m – MATLAB binary data file, which contains magnetic field maps for all points of averaged RR interval.
4. Vectors.m – MATLAB binary data file, which contains currents vector maps for all points of averaged interval.

External software requirements: CarMag Package Shell and Currents software item.

IX. Database

Database (DB) of MCG examinations was developed to store the relevant information of patients, MCG measurements and examination results. DB includes cards for patients with the options of fast-access search, selection and sorting. An individual MCG examinations table is subloaded for every patient. There is a table of final results for every examination. It is provided the input of detailed comments for every patient as well as for every examination or result. The operator can edit or delete in DB any previously stored information.

Computer Database includes many files in Sub-directory BASE and manages by Package Shell with help of Borland Database Engine (BDE). The BDE realizes recording, storage, updating and extracting of raw MCG data by means of plane-structured compressed binary files. Each compressed file of this type corresponds to a single complete MCG examination and contains the files of raw MCG records, major intermediate and final results.

Three operation modes are provided for DB of MCG examinations:

1) Filling out of a card for a new patient and (or) examination. Initialization of the work of data input program module. Storage of the input data in DB.

2) Extraction of the previously accumulated MCG data from DB for further processing, analysis and interpretation.

3) Storage of analysis results for processed MCG examination in DB.

As extra options, it is realized creation of a new blank DB, copying of the current DB to archive and opening another available DB. Based on BDE we can create and manage our Databases.

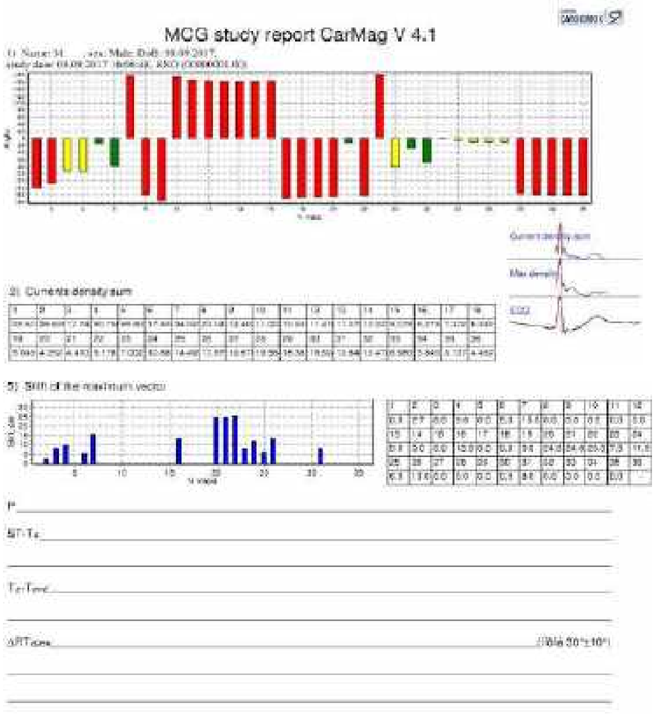
X. Current Density Reconstruction

The software module CURRENTS is intended to solve inverse problem and to calculate Current Density Vectors (CDV) maps from averaged MCG maps. This module is activated by TOPOLOGY item. External software requirements for this module: Topology software item and Matlab Runtime Compiler.

Used files: 1) Input data files *Fields.m* – MATLAB binary data file, which contains magnetic field maps for all time-points of averaged RR interval. 2) Output data files *Vectors.m* – MATLAB binary data file, which contains CDV maps for all time-points of averaged RR interval.

Essence of the module is calculation of images to visualize the heart electric activity in view of grid of 10x10 current vectors distributed within area 20X20 cm into frontal plane, based on pseudo-current approach described early at [10].

In result of data processing of each record from database it is created final report with quantitative parameters and a few pages of CDV maps. The report and 1 page of CDV maps is shown at Fig. 12-13.



MCG study report CarMag V 4.1

1) Name: M... sex: Male, DoB: 08.09.2017,
study date: 08.09.2017 16:06:48, RNO /0000000001.00/

2) Currents density sum

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

3) Shift of the maximum vector

1	2	3	4	5	6	7	8	9	10	11	12
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
13	14	15	16	17	18	19	20	21	22	23	24
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
25	26	27	28	29	30	31	32	33	34	35	36
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

P ______________________________

ST-Te ______________________________

____

Te-Tend ______________________________

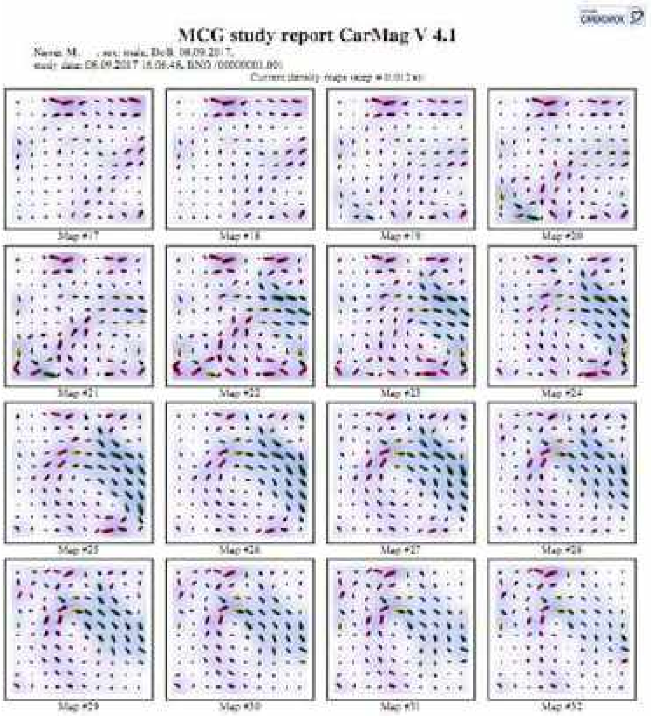
____

ΔRTαes ______________________________ (Nla 50°±10°)

____

____

Fig. 12. Final report with quantitative parameters



MCG study report CarMag V 4.1

Name: M... sex: male, DoB: 08.09.2017,
study date: 08.09.2017 16:06:48, RNO /0000000001.00/

Current density maps (step #0.012 s)

Fig. 13. Final report. One of pages with CDV maps

Conclusions

Description of developed software package CARMAG showed in detail. It was presented structure schemes of full package, embedded software, acquisition module, CARMAG package shell, Topology module, and Currents module. The functioning of MCG examinations database was described. In addition, figures of general view of software main window, Input9 and Topology modules, final report of software with quantitative parameters and Current Density Vectors maps are presented.

Acknowledgments

The software has been developed within the framework of STCU (Science and Technology Center in Ukraine) projects “Development of method and apparatus for 3D imaging of electrophysiological activity of biological tissues” #P624 (2014-2016), #P624a (2016-2017), #P624b (2017-2018), #P624c (2018-2023), partner company Oxford Cardiomox Ltd, UK. Authors from Glushkov Institute of Cybernetics give thanks for financial support to Suzhou Cardiomox Ltd (Suzhou, Jiangsu, P.R.China) and personally to Company CEO Xie Feng. Also authors give thanks for mutual collaboration to Dr Mykhailo Primin, Ilya Chaikovskyi, researcher Ihor Nedaivoda from Glushkov Institute of Cybernetics, and Tao Miao, Shouzhang Zhu from Suzhou Cardiomox Ltd.

References

[1] Draft National Strategy for the Development of New Health Care System in Ukraine for the period 2015-2025 (in Ukrainian). - URL: http://www.apteka.ua/artikle/315522.

[2] S.Ebmeyer, I.Chaikovsky, B.Hailer, R.Erbel, H.Wojczik, M.Budnyk, R.Simon, “Predictive value of the magnetocardiogram for location of regional ischemia or infarction as detected by quantitative analysis of the coronary arteriogram”, International Congress Series, 1300, 2007, pp. 463-467. DOI: 10.1016/j.ics.2007.02.008.

[3] A.S. Dovbysh, S.S. Martynenko, A.S. Kovalenko, N.N. Budnyk, “Information-extreme algorithm for recognizing current distribution maps in magnetocardiography”, Journal of Automation and

Information Sciences, 43 (2), 2011, pp. 63-70. DOI: 10.1615/ JAutomatInfScien.v43.i2.60.

[4] I. Chaykovskyy, M. Najafian, M. Budnyk, S. Martynenko, A. Dovbysh, O. Kovalenko, "Development of pattern recognition method for diagnosis of myocardial ischemia and noncoronarogenic myocardial diseases based on current density distribution maps", In: 17th International Conference on Biomagnetism: Advances in Biomagnetism – Biomag2010. IFMBE Proceedings, Supek, S., Sušac, A. (eds), vol 28. Springer, Berlin, Heidelberg. 2010, pp. 424-427. https://doi.org/10.1007/978-3-642-12197-5_101.

[5] I. Voytovych, M. Budnyk, Y. Minov, L. Artemenko, E. Paslavsky, P.Shpilyovy, "First Ukrainian multi-channel magnetocardiograph: Assembling and testing", Proceedings of the 2nd IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2003, art. no. 1249507, 2003, pp. 16-19. DOI: 10.1109/IDAACS.2003.1249507.

[6] I.Voytovych, M.Primin, V.Vasyliev, P.Sutkovyy, M.Budnyk, I.Nedayvoda, A.Rusanov, T.Ryzhenko, "Multi-channel magneto-cardiograph: Control and software", Proceedings of the 2nd IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS'2003, art. no. 1249591, 2003, pp. 382-384. DOI: 10.1109/IDAACS.2003.1249591.

[7] M. Budnyk, V. Sosnytskyi, I. Voitovych, V. Maiko, Y. Minov, P.Sutkovyi, O. Zakorchenyi, T. Ryzhenko, V. Budnyk, "Development of 4-channel cardiomagnetic scaner and technical requirements for 9-channel scaner to diagnose the heart abnormalities", Proceedings of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS'2011, 1, art. no. 6072717, 2011, pp. 91-96. DOI: 10.1109/IDAACS.2011.6072717.

[8] V. Budnyk, M. Budnyk, V. Sosnytskyy, X. Feng, T. Miao, S. Zhu, P. Sutkovyi, Y. Minov, P. Spylovyy, M. Mudrenko, Y. Melnyk, W. Ji. "Development of 9-channel Magnetocardiograph CARDIOMOX MCG-9" // Conference materials of IEEE XIIIth International Conference on Electronics and Information Technologies (ELIT), 2023. (in this issue).

[9] Renesas. [Online]. Availeble: https://www.renesas.com/eu/en.

[10] M. Budnyk, Yu. Frolov, A. Moroz, "Creating software for reconstruction of the heart electric activity within the framework of pseudo-current approach", Proc. Summer School "Biological & Medical Informatics & Cybernetics" (BMIC), Kyiv: Glushkov Institute of Cybernetics, Eds. M. Budnyk, I. Voitovych, 2012, Part 1. pp. 65-70 (in Ukrainian).

Study of the Morphology of Macroporous Silicon Obtained by Metal-Assisted Etching Using Chromium

Mykola S. Kukurudziak
Engineering production complex 1
Rhythm Optoelectronics Shareholding Company
Chernivtsi, Ukraine
mykola.kukurudzyak@gmail.com

***Abstract* — The article examines the morphology of porous silicon doped with phosphorus and boron obtained by metal-stimulated etching using Cr. Selective Sirtles etchant with the composition of HF - 100 cm^3, CrO$_3$ - 50 g, H$_2$O - 120 cm^3 was chosen. The application of the Cr film was carried out by thermal sputtering in a vacuum. The thickness of the film was 10-40 nm. After the etching process, a macroporous structure with pyramid-shaped pores was revealed on the surface of the plates. It was established that with an increase in the thickness of the metal catalyst film and with an increase in the duration of etching, the size of the pores and etching pits increases. A uniform macroporous structure can be obtained by etching silicon with a chromium film 10-15 nm thick. When using a film with a thickness of 20-30 nm, the size of etching pits increases sharply. When using a chromium film with a thickness of more than 30 nm, uneven etching of the film is possible, which leads to uneven pore formation. The size of the pores on the surface of silicon doped with boron is much smaller than on the surface of silicon doped with phosphorus, which is caused by a decrease in the concentration of the doping impurity.**

Keywords* — *metal-assisted etching, selective etchant, porous silicon.

I. Introduction

With the development of science and technology, scientists are increasingly interested in low-dimensional structures, their properties and manufacturing processes. One of the most common objects with low-dimensional morphology is porous silicon (por-Si). Due to the possibility of creating porous structures with given optical properties, por-Si is used in solar energy as anti-reflective textured coatings [1]. In medicine, por-Si is used as antiviral adsorption nanoparticles [2]. In the technology of silicon integrated circuits or photodetectors, por-Si can be used for gettering of generation-recombination centers with a broken layer [3]. por-Si is also actively used for the production of membranes capable of separating molecules by size [4], etc.

There are many methods of obtaining a por-Si. The reactive ion etching method is often used in the technology of microelectronic processes. This technology allows obtaining ordered porous structures with controlled parameters, but requires complex technological implementation [5]. Plasma chemical etching methods are also known, but the most common are electrochemical and chemical etching of monocrystalline Si plates, in particular metal-stimulated etching [6]. The most commonly used elements for metal-stimulated etching are Au and Ag due to their high manufacturability. Metals can be deposited on the Si substrate using vacuum sputtering processes or chemical deposition methods from solutions. Vacuum sputtering allows for the formation of ordered structures, while the chemical deposition method is simpler and cheaper and is used when the morphology of the final substrate is not important. Note that these metals are expensive, the use of which in the production of electronics significantly increases the price of products. Accordingly, the current scientific and technical task is the search for new methods and materials that can be used to obtain textured materials.

We established the possibility of por-Si formation by means of metal-stimulated etching with Chromium. When studying the above, some heterogeneity of the formation of pores caused by various factors was seen. This phenomenon required additional research to establish the mechanisms of influence on the uniformity of pore formation. When reviewing scientific sources, it is seen that many works are devoted to the issue of por-Si formation by metal-stimulated etching with the help of Au. In particular, this method was proposed back in 2000 by Lee and Bon [7], and they also presented etching in a HF/H$_2$O$_2$ solution using Ag, Au, and Pt. In [6, 8], it was investigated that Ag and Au nanoparticles formed in solution lead to the formation of straight pores during etching, while straight or helical pores can be obtained with the help of a Pt catalyst. And in [9], the authors reported that Pt nanoparticles move chaotically during etching, which leads to curved pores without a uniform etching direction. Also, [10] reports the possibility of forming pyramidal pores by chemical etching of Si with Ni. However, no information has been found on metal-stimulated etching of silicon using chromium films. Therefore, the purpose of this work is to study the method of metal-stimulated texturing of silicon with the help of chromium, and to study the morphology of the samples obtained by this method.

II. Experimental Details

Monocrystalline *p*-type silicon of the orientation [111] with a resistivity $\rho\approx18$ kΩ·cm was used for the experiments, which corresponds to the concentration of acceptors $N_A\approx7.7\cdot10^{11}$ cm^{-3}. In order to study the structure of por-Si on the surface of doped silicon of different conductivity types, diffusion of phosphorus and boron from planar solid-state sources was carried out thermally according to the methods given in [11] and [12], respectively. For this, the substrates were pre-oxidized, photolithography was carried out to obtain an arbitrary topology, and actual diffusion was carried out. The concentration of applied phosphorus was N_{P0}=4.3-4.7·10^{20} cm^{-3} (R_S=2.1-2.7 Ω/□). The concentration of

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

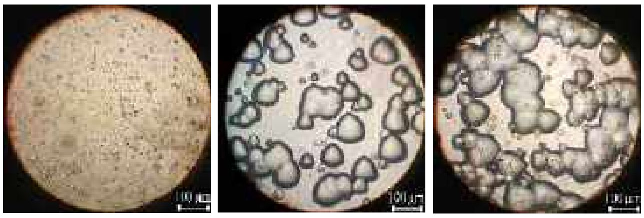
introduced boron was N_{B0}=2.9-3.9·10^{20} cm^{-3} (R_S=16-25 Ω/□). Next, the Cr film was deposited by thermal spraying in a vacuum according to the method given in [13]. The thickness of the films reached d_{Cr}=10-40 nm.

Further, after chemical treatment in a boiling Caro mixture and an ammonia-peroxide solution, the surface of the silicon substrates was etched to form por-Si. Selective Sirtle's etchant with the composition of HF - 100 cm^3, CrO_3 - 50 g, H_2O - 120 cm^3 was chosen. Metal-stimulated etching of silicon was carried out in a static solution at room temperature for 1-10 minutes. The obtained structures were studied in microscopes with different magnifications and with the help of an atomic force microscope (AFM) NT-206.

III. Result and Discussion of the Research

A. *Metal-assisted etching of p-Si*

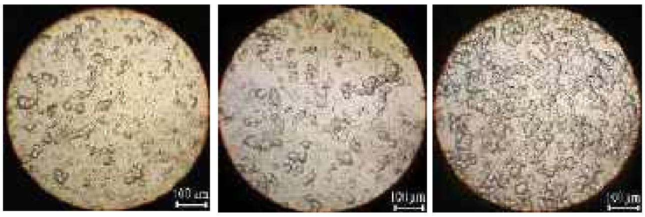
After metal-assisted etching using chromium quasi-pyramidal pits of etching, which were macropores, were observed on the surface of the substrates. When the duration of etching was increased, a significant increase in the size of the etching pits was observed (Fig. 1).



a) *b)* *c)*

Fig. 1. Macroporous structure on the p-Si surface at different etching durations (d_{Cr}=20 nm): a) t=1 min; b) t=5 minutes; c) t=7 min.

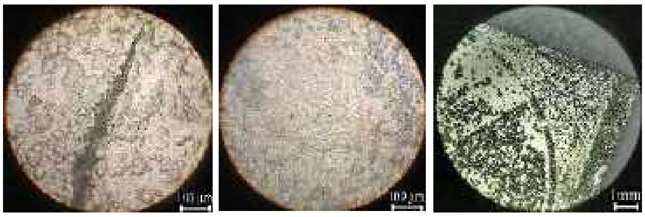
It was established that the density of pores increases with an increase in the thickness of the film (Fig. 2). Note that with an increase in the duration of etching or the thickness of Cr, etching pits can combine into complexes (Fig. 1, 2).



a) *b)* *c)*

Fig. 2. Image of the structure of the pores formed during etching for 3 min at different thicknesses of the Cr film: a) d_{Cr}=10 nm; b) d_{Cr}=20 nm; c) d_{Cr}=30 nm.

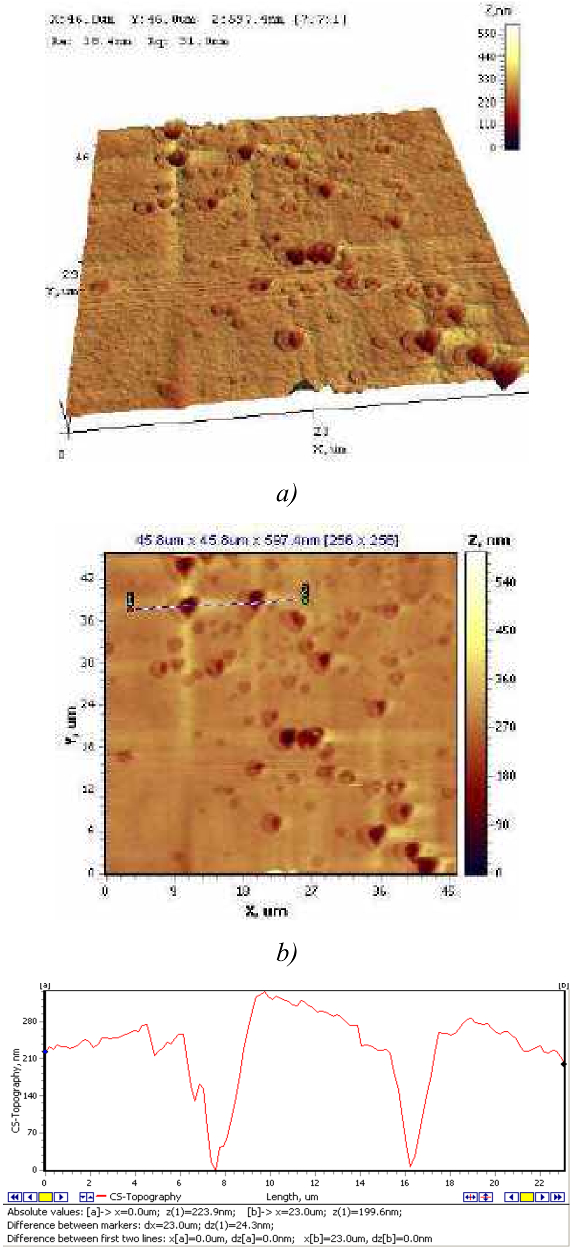
Also, the density of pores increases in the areas of presence of defects or in areas with increased mechanical stresses, in particular on the periphery of the plates, because after the operation of cutting the plates into crystals, abundant defect formation is possible along the cutting line (Fig. 3).



a) *b)* *c)*

Fig. 3. Image of the porous structure of silicon along the scratch (a) and on the periphery of the crystal (b, c).

For a detailed analysis of the morphology of the porous *p*-Si surface, an AFM image was obtained (Fig. 4).



Absolute values: [a]-> x=0.0um; z(1)=223.9nm; [b]-> x=23.0um; z(1)=199.6nm;
Difference between markers: dx=23.0um; dz(1)=24.3nm;
Difference between first two lines: x[a]=0.0um, dz[a]=0.0nm; x[b]=23.0um, dz[b]=0.0nm

a)

b)

c)

Fig. 4. AFM image of porous p-Si after etching for 1 min: a) 3D image; b) 2D image; c) structure profile.

From Fig. 4, it can be seen that etching for 1 min allows to obtain pores with a depth of up to 300 nm and a diameter of up to 2 µm.

B. *Metal-assisted etching of n^+-Si*

Metal-stimulated etching of phosphorus-doped silicon with a chromium thickness of 20-30 nm and different process durations was carried out. Quasi-hexagonal etching pits of different sizes were found on the surface of the plates after this chemical treatment (Fig. 5). The hexagonal shape of the etching pits is explained by the orientation of silicon, although pits in the form of equilateral triangles or pyramidal depressions are characteristic for the [111] crystallographic orientation during anisotropic etching, but with a significant increase in the duration of selective processing, the specified structures are modified and hexagonal pits are formed.

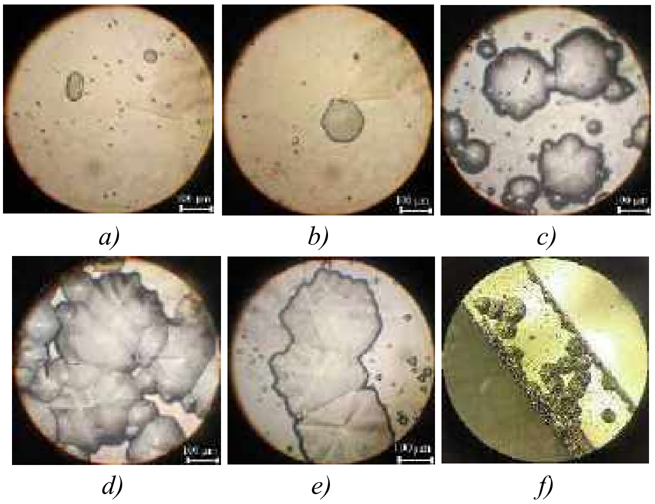


Fig. 5. Image of a silicon wafer after metal-stimulated etching with duration (d_{Cr}=20-30 nm) : a) 1 min; b) 3 min; c) 5 min; d) 7 min; e)-f) 10 min.

From Fig. 1, it can be seen that the size of the etching pits increases when the etching time increases. When the duration of chemical treatment is *t*=1 min, individual pits up to 60 µm in size are observed (Fig. 5a). At *t*=3 min, single etching pits up to 100 µm in size are also observed, which begin to acquire a hexagonal shape. It should be noted that after etching for 1-3 minutes, the remains of a chromium film are observed on the surface of the plates, which does not have time to dissolve in the etchant due to the short duration of the process (Fig. 5a, porb). At *t*=5 min, the size of pits reached up to 120 µm, and at *t*=7 min - up to 150-200 µm (Fig. 5c, d). As the duration of metal-stimulated etching increased, the density of pits increased and their merging into complexes occurred.

When the duration of the chemical treatment is increased for more than 7 minutes, the surface of the substrate is completely covered with combined etching pits. In this case, consideration of individual digestion objects is impossible. Therefore, in order to study the appearance of etching pits of a higher etching treatment, metal-stimulated etching was carried out with the help of a silicon oxide film grown by the method of wet oxidation [14]. Since with this method of oxidation a film with pores is formed, it is possible to study local objects of etching when using it. In this case, the chromium film was applied directly to the silicon oxide. Etching on the silicon surface lasting 10 min was carried out in this way (Fig. 5e, f). In this case, the etching pits were up to 250 µm in size. There was no further increase in the duration of the process. It should be noted that hydrofluoric acid itself would have completely etched the silicon oxide at this process duration, but given that the etchant is a mixture, the oxide etching is much slower, which allows us to describe the structuring.

It should be noted that when using a chromium film with a thickness of more than 30 microns, its poor etching is possible, which leads to non-uniform etching, uneven formation of pores and the presence of areas with remnants of the metal film (Fig. 6).

It is also worth noting that during metal-assisted etching of substrates with a high density of dislocations (usually at high concentrations of diffusant), there is a near-surface growth of the areas where dislocations exit to the surface and an increase in their etching pits due to the course of the selective etching reaction becoming brighter. These dislocations disrupt the surface structure of textured silicon and may represent the centers of pits-"craters" created during etching (Fig. 7).

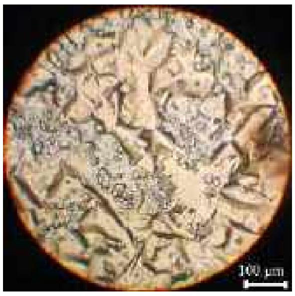


Fig. 6. Image of the surface of a silicon wafer after metal-stimulated etching with d_{Cr}=35-40 nm.

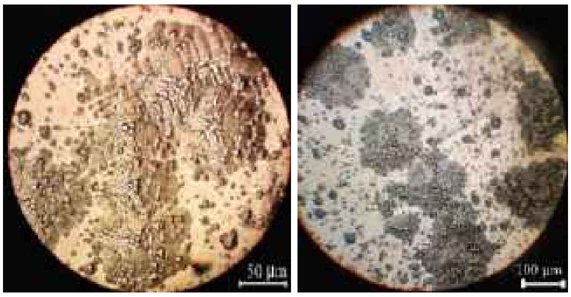


Fig. 7. Surface of textured silicon with an increased density of dislocations (t=5 min, d_{Cr}=20-30 nm).

The obtained textures on the surface of n^+-Si are interesting only from the precise point of view of crystallography and fundamental science. For the possibility of using porous silicon in solar energy or other fields of electronics, the size of the pores should be much smaller. To obtain pores of reduced size, it is worth using chromium films with a thickness of less than 10 nm. A metal-assisted etching process was carried out with a chromium thickness of about 10 nm for 5 (Fig. 8a) and 10 (Fig. 8b) min.

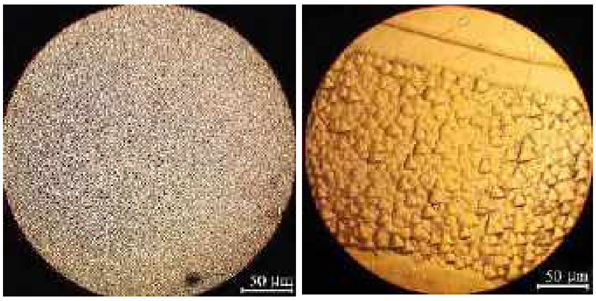
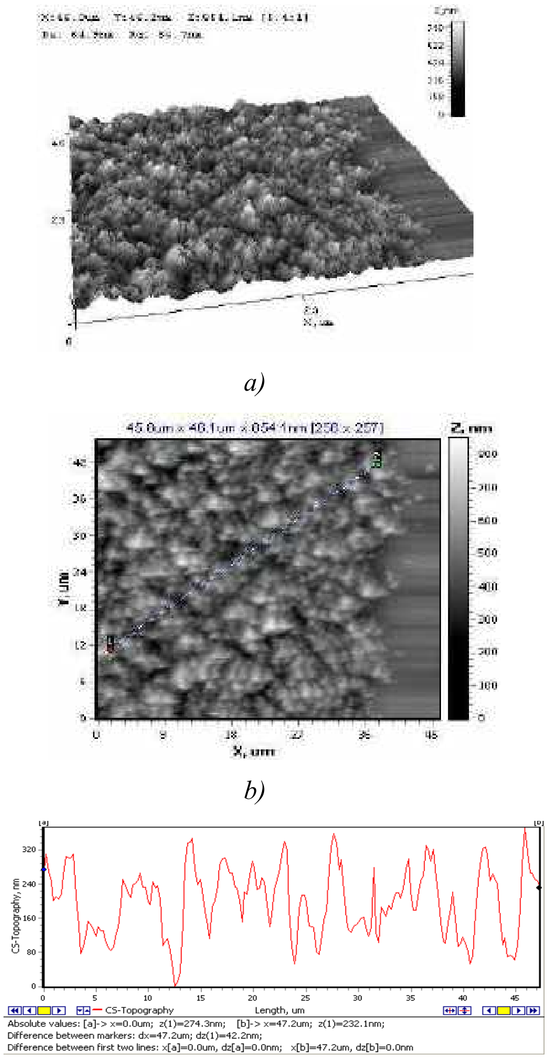


Fig. 8. Image of porous n^+-Si obtained by etching with $d_{Cr}\approx$10 nm and t=5 min (a) and t=10 min (b).

From Fig. 8, it is possible to see a homogeneous macroporous structure, which is significantly different from the structures obtained with the use of a greater thickness of the chromium film. For a detailed assessment of the shape and size of the pores, an AFM image of the textured surface was obtained (Fig. 9 and Fig. 10). On Fig. 9 shows the surface of porous silicon obtained by etching for 5 min. In this case, pyramid-shaped pores with a depth of up to 300 nm and a width of up to 2 µm were obtained. When the etching time was increased to 10 min, the pore depth reached 800 nm and the width reached 7 µm (Fig. 10).

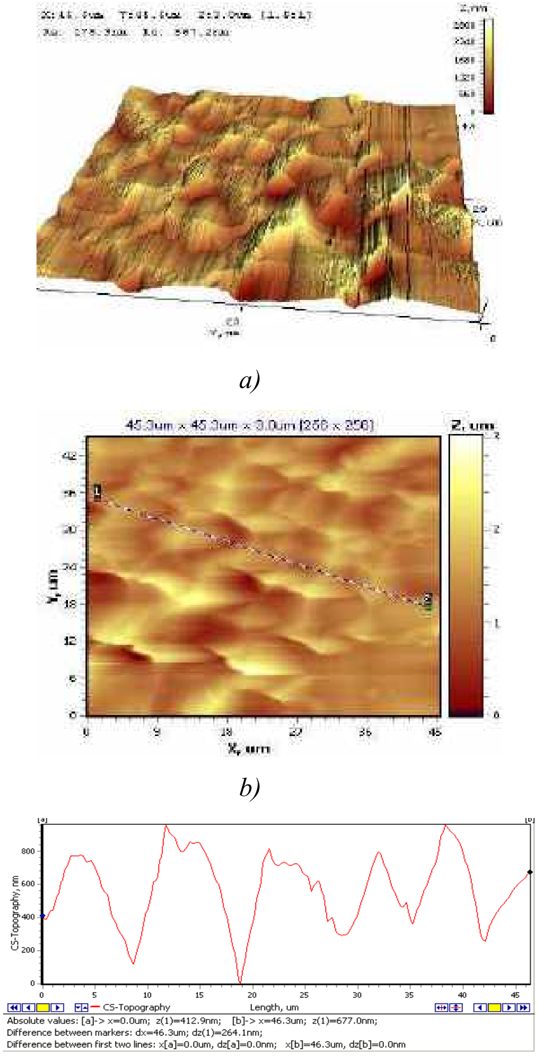


a)

b)

Absolute values: [a]-> x=0.0um; z(1)=274.3nm; [b]-> x=47.2um; z(1)=232.1nm;
Difference between markers: dx=47.2um; dz(1)=42.2nm;
Difference between first two lines: x[a]=0.0um, dz[a]=0.0nm; x[b]=47.2um, dz[b]=0.0nm

c)



a)

b)

Absolute values: [a]-> x=0.0um; z(1)=412.9nm; [b]-> x=46.3um; z(1)=677.0nm;
Difference between markers: dx=46.3um; dz(1)=264.1nm;
Difference between first two lines: x[a]=0.0um, dz[a]=0.0nm; x[b]=46.3um, dz[b]=0.0nm

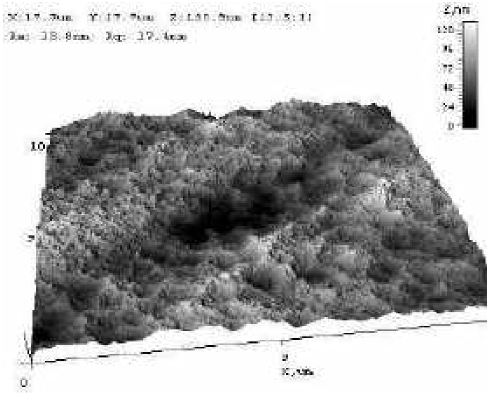
c)

Fig. 9. AFM image of porous n$^+$-Si after etching for 5 min (d_{Cr}≈10 nm): a) 3D image; b) 2D image; c) structure profile.

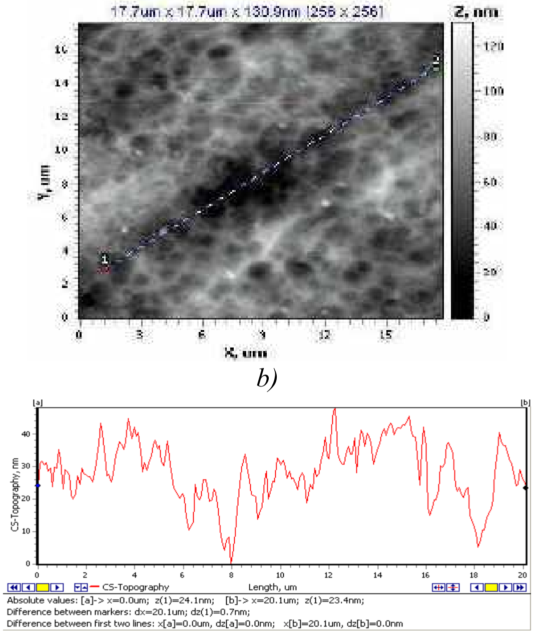
Fig. 10. AFM image of porous n$^+$-Si after etching for 10 min (d_{Cr}≈10 nm): a) 3D image; b) 2D image; c) structure profile.

C. *Metal-assisted etching of p$^+$-Si*

A porous structure was also obtained on the *p*$^+$-Si surface at *d*=10 nm and *t*=5 min. In this deposit, the pore size was significantly smaller than when using the same regimes for *n*$^+$-Si due to the difference in the concentrations of alloying impurities [15]. In this case, the pore depth reached 5-30 nm and the width 0.1-1 μm (Fig. 11).



b)



Absolute values: [a]-> x=0.0um; z(1)=24.1nm; [b]-> x=20.1um; z(1)=23.4nm;
Difference between markers: dx=20.1um; dz(1)=0.7nm;
Difference between first two lines: x[a]=0.0um, dz[a]=0.0nm; x[b]=20.1um, dz[b]=0.0nm

b)

c)

Fig. 11. AFM image of porous p$^+$-Si after etching for 10 min (d_{Cr}≈10 nm): a) 3D image; b) 2D image; c) structure profile.

The resulting homogeneous macroporous silicon surfaces of various conductivity types can be used as anti-reflective coatings, for gaitering of generation-recombination centers, etc.

Conclusions

The morphology of macroporous silicon doped with phosphorus and boron was investigated. The following conclusions were made:

1. After metal-assisted etching using chromium quasi-pyramidal pits of etching, which were macropores, were observed on the surface of the substrates.

2. When the duration of etching was increased, a significant increase in the size of the etching pits was observed.

3. The density of pores increases with an increase in the thickness of the film.

4. The density of pores increases in the areas of presence of defects or in areas with increased mechanical stresses.

5. With an increase in the duration of etching, the pore size increases.

6. When the duration of the chemical treatment is increased for more than 7 minutes, the surface of the substrate is completely covered with combined etching pits with size of 150-200 μm.

7. When using a chromium film with a thickness of more than 30 microns, its poor etching is possible, which leads to non-uniform etching, uneven formation of pores and the presence of areas with remnants of the metal film.

8. To obtain pores of reduced size, chromium films with a thickness of less than 10 nm should be used. After 5 minutes of etching, pyramidal pores with a depth of up to 300 nm and a width of up to 2 μm were obtained. With an increase in the etching time to 10 minutes, the pore depth reached 800 nm and the width was 7 μm.

9. A porous structure was also obtained on the p^+-Si surface at a chromium film thickness of 10 nm and an etching time of 5 min. In this case, the pore depth reached 5-30 nm, and the width was 0.1-1 μm.

References

[1] Mouafki, A. M., Bouaïcha, F., Hedibi, A., & Gueddim, A. (2022). Porous Silicon Antireflective Coatings for Silicon Solar Cells. Engineering, Technology & Applied Science Research, 12(2), 8354-8358. DOI: https://doi.org/10.48084/etasr.4803

[2] Osminkina, L. A., Agafilushkina, S. N., Kropotkina, E. A., Saushkin, N. Y., Bozhev, I. V., Abramchuk, S. S., ... & Gambaryan, A. S. (2022). Antiviral adsorption activity of porous silicon nanoparticles against different pathogenic human viruses. Bioactive materials, 7, 39-46. DOI: https://doi.org/10.1016/j.bioactmat.2021.06.001

[3] Pilipenko, V. A., Gorushko, V. A., Petlitsky, A. N., Ponaryadov, V. V., Turtsevich, A. S., & Shvedov, S. V. (2013). Methods and mechanisms of hetterization of silicon structures in integrated circuit manufacturing. Technology and design in electronic equipment, 2-3, 43-57.

[4] Vercauteren, R., Scheen, G., Raskin, J. P., & Francis, L. A. (2021). Porous silicon membranes and their applications: Recent advances. Sensors and Actuators A: Physical, 318, 112486. DOI: https://doi.org/10.1016/j.sna.2020.112486

[5] Zulkifli, F., Radzali, R., Abd Rahim, A.F., Mahmood, A., Mohd Razali, N.S. and Abu Bakar, A. (2022), "Influence of different etching methods on the structural properties of porous silicon", Microelectronics International, vol. 39, no. 3, pp. 101-109. DOI: https://doi.org/10.1108/MI-01-2022-0009

[6] A. A. Leonardi, M. J. lo Faro, and A. Irrera, "Silicon Nanowires Synthesis by Metal-Assisted Chemical Etching: A Review," Nanomaterials, vol. 11, no. 2, p. 383, Feb. 2021, DOI: https://doi.org/10.3390/nano11020383.

[7] Li, X., & Bohn, P. W. (2000). Metal-assisted chemical etching in HF/H 2 O 2 produces porous silicon. Applied Physics Letters, 77(16), 2572-2574. Doi: https://doi.org/10.1063/1.1319191

[8] Salem, A.M.S.; Harraz, F.A.; El-Sheikh, S.M.; Ismat Shah, S. Novel Si nanostructures via Ag-assisted chemical etching route on single and polycrystalline substrates. Mater. Sci. Eng. B Solid-State Mater. Adv. Technol. 2020, 262, 114793. Doi: https://doi.org/10.1016/j.mseb.2020.114793

[9] Chen, C.Y.; Wu, C.S.; Chou, C.J.; Yen, T.J. Morphological control of single-crystalline silicon nanowire arrays near room temperature. Adv. Mater. 2008, 20, 3811–3815. DOI: https://doi.org/10.1002/adma.200702788

[10] Yue, Z., Shen, H., Jiang, Y. et al. Novel and low reflective silicon surface fabricated by Ni-assisted electroless etching and coated with atomic layer deposited Al_2O_3 film. Appl. Phys. A 114, 813–817 (2014). DOI: https://doi.org/10.1007/s00339-013-7670-y

[11] Kukurudziak, M. S. (2022). Diffusion of phosphorus in technology for manufacturing silicon pin photodiodes. Semiconductor Physics, Quantum Electronics & Optoelectronics, 25(4), 385-393. https://doi.org/10.15407/spqeo25.04.385

[12] M. S. Kukurudziak and E. V. Maistruk, "Features of Diffusion Doping and Boron Gettering of Silicon p-i-n Photodiodes," 2022 IEEE 3rd KhPI Week on Advanced Technology (KhPIWeek), 2022, pp. 1-6, doi: https://doi.org/10.1109/KhPIWeek57572.2022.9916420

[13] M. S. Kukurudziak, E. V. Maistruk, "Influence of chromium sublayer on silicon P-I-N photodiodes responsivity," Proc. SPIE 12126, Fifteenth International Conference on Correlation Optics, 121261V (20 December 2021); DOI: https://doi.org/10.1117/12.2616170

[14] Kukurudziak, M. S. (2023). Problems of Masking and Anti-Reflective SiO_2 in Silicon Technology. East European Journal of Physics, (2), 289-295. https://doi.org/10.26565/2312-4334-2023-2-33

[15] Ouyang, H., Christophersen, M., & Fauchet, P. M. (2005). Enhanced control of porous silicon morphology from macropore to mesopore formation. physica status solidi (a), 202(8), 1396-1401. DOI: https://doi.org/10.1002/pssa.200461112

Degradation of Silicon Resistivity During Thermal Operations in *p-i-n* Photodiodes Technology

Mykola S. Kukurudziak
Engineering production complex 1
Rhythm Optoelectronics Shareholding Company
Chernivtsi, Ukraine
mykola.kukurudzyak@gmail.com

Abstract — **The article investigates the degradation of silicon resistivity in the process of manufacturing silicon four-element p-i-n photodiodes with a guard ring. It was established that the main reason for the deterioration of the electrophysical parameters of silicon is high-temperature processing and actual thermal shocks. To minimize the deterioration of material characteristics, it is worth using technology with a minimum of thermal operations and using thermal operations with the lowest possible temperature. The degree of degradation of the resistivity of silicon when using planar technology using two-stage diffusion of phosphorus from solid sourcesor or using one-stage diffusion of phosphorus from liquid PCl$_3$ diffusant and mesa-technology was investigated. A non-destructive method of controlling the resistivity of finished crystals by determining the reverse bias voltage, at which the space charge region acquires the width of the substrate, is proposed. It was established that the operations of oxidation and driving-in of phosphorus will reduce the resistivity of silicon by 1-1.5 kΩ·cm and 2 kΩ·cm, respectively, and during the operation of boron gettering of the reverse side of substrate, some restoration of the resistivity occurs. It was also established that when using mesa-technology technology, it is possible to actually avoid the degradation of the electrical resistance due to the exclusion of the high-temperature oxidation operation.**

Keywords* — *resistivity, silicon, photodiode, dark current, responsivity.

I. Introduction

Silicon was and remains the main material of solid-state electronics due to its unique properties, high manufacturability [1] and sufficient reserves of natural raw materials [2]. The electrophysical parameters of silicon depend on the methods of obtaining it and subsequent technological operations in the process of manufacturing electronic elements. The main parameters of Si are resistivity (ρ), lifetime (τ), concentration and mobility of charge carriers. Real silicon crystals usually have structural defects formed during ingot growth. These can be growth dislocations, point defects (vacancies, inclusions), twins, packing defects. With the development of electronics and methods of obtaining semiconductor materials, the possibility of growing dislocation-free single-crystal Si has been achieved, which excludes the possible presence of the above-mentioned defects [3]. But unlike growth defects, there are also technologically introduced defects: dislocations formed during thermal operations, oxidation defects of packaging, impurities of uncontrolled metal impurities, etc. [4] These defects negatively affect the parameters of the base material, which must be taken into account when modeling or calculating the final parameters of devices. Avoiding the degradation of electrophysical parameters of semiconductor material in the technological process is an important and urgent scientific and technical task.

During our production of silicon *p-i-n* photodiodes (PD), some degradation of the resistivity of the material was observed after high-temperature operations. This phenomenon required a detailed study to establish the causes of its occurrence and methods of their avoidance.

When reviewing literary sources, it was seen that many works are devoted to the issue of degradation of electrophysical parameters of semiconductor materials. In [5] shown that the gas phase is a potential source for contamination of silicon which can cause a degradation of the charge carrier lifetime as the back diffusion of segregated metals at the surface of plates. Beside Fe, significant amounts of Co and Ni are introduced which diffuse deeper into the material, affecting the material quality. The results clearly show that a contamination is introduced into the silicon, originating both from the furnace and the purging gas. This contamination is highest at a low furnace pressure and a low at high purging rate. In [6] it is indicated that avoiding contamination of silicon for solar cells during high-temperature processing steps is a key issue. Here, it is shown that interactions with the gas phase also are a potential source of contamination. Thermodynamic calculations performed for a temperature range of 373 K to 1873 K (100 °C to 1600 °C) and total pressure of 10 kPa predict the formation of volatile species that are harmful for photovoltaic properties. In [7] was found that after the high-temperature processing (1200°C) of transmutation doped *n*-type silicon crystals for 2 h with the cooling rate of 15°C/min, and also for 72 h and with all the cooling rates studied (1, 15 and 1000°C/min), the generation of deep donor centers occurs in their volume. It was also established that the rate of cooling after annealing also significantly affects the degradation of material parameters. It is also known [8] that thermal shocks during heat treatment contribute to the formation of defects in substrates and the degradation of ρ and τ.

From the above, it can be concluded that the main reason for the degradation of the electrophysical parameters of silicon is high-temperature processing and actual thermal shocks. In order to minimize the deterioration of material characteristics, it is worth using technology with a minimum of thermal operations and using thermal operations with the lowest possible temperature. In planar technology, the use of mesa-structures [9-11] and one-stage diffusion methods, as opposed to two-stage ones [12], allows to reduce the number of thermal operations. Therefore, the purpose of this work is to investigate the degree of degradation of the resistivity of silicon in *p-i-n* PDs technology depending on the combination of thermal operations, as well as the search for non-destructive methods of determining the resistivity of the base material of finished products.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

II. Experimental Details

The research was carried out on silicon four-element *p-i-n* PDs with a guard ring (Fig. 1) for work at a wavelength of Fifvariants of the technology: by diffusion-planar using two-stage diffusion of phosphorus from planar sources (PD$_1$), and by mesotechnology using two-stage diffusion of phosphorus from planar sources (PD$_2$), and by diffusion-planar using one-stage diffusion of phosphorus using liquid diffusant PCl$_3$ (PD$_3$). After manufacturing, the final parameters of the PDs were compared. Dislocation-free *p*-type single crystal silicon (*Cz*-Si) with [111] orientation, resistivity ρ≈20-24 kΩ·cm and life time of minor charge carriers τ≈1.8-2.2 ms was used. The thickness of the substrates reached x≈440 µm.

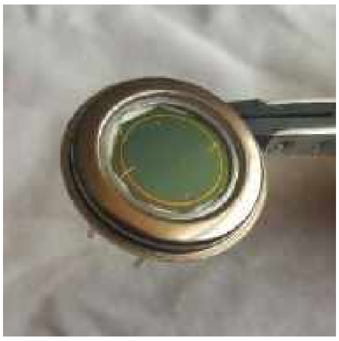
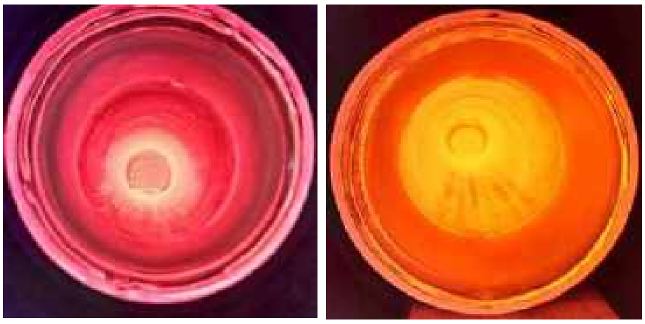


Fig. 1. Silicon four-element *p-i-n* PD with a guard ring

The PD$_1$ technological route consisted of a complex of four thermal operations and three photolithographies according to the regimes given in [12]: semiconductor substrates were oxidized according to the principle of dry-wet-dry oxidation at a temperature of T=1423 K; photolithography was carried out to create windows for phosphorus diffusion; diffusion of phosphorus (predeposition) to the front side to create n^+-type responsive elements (RE) and guard ring at a temperature of T=1323 K (Fig. 2a); drive-in of phosphorus to redistribute the alloying impurity and increase the depth of the n^+-p-junction (x_{n+-p}=4-5 µm) at a temperature of T=1423 K; diffusion of boron to the reverse side of the substrate to create a p^+-type ohmic contact (x_{p+-p}=1-2 µm) at a temperature of T=1223 K (Fig. 2b); photolithography for creating contact windows; sputtering of Cr-Au on the front and back sides.



a) b)

Fig. 2. Silicon wafers in a quartz reactor at T=1223 K (a) and T=1323 K (b).

The PD$_2$ technological route consisted of a complex of three thermal operations and three photolithographies [9]. It differed from PD$_1$ in that the first thermal operation was the predeposition of phosphorus from planar sources, and the formation of the crystal topology was carried out by etching the meso-profile using the chemical dynamic polishing method [13] after photolithography. All subsequent thermal operations were carried out as in the case of PD$_1$. This made it possible to exclude a high-temperature oxidation operation.

The PD$_3$ technological route consisted of a complex of three thermal operations and three photolithographies. It differed from PD$_1$ in that phosphorus diffusion was carried out using PCl$_3$ during one thermal operation in an oxidizing medium at a temperature of T=1323 K according to the regimes given in [14]. This made it possible to exclude the high-temperature operation of drive-in of phosphorus. All other thermal operations were carried out as in the case of PD$_1$.

The surface concentration of phosphorus in all versions of the technology reached N_{P0}=4.1-4.15·10^{20}cm^{-2} (R_s≈2.7 Ω /□), boron – N_{B0}=4.3-4.35·10^{20}cm^{-2} (R_s ≈18 Ω /□).

The resistivity of the samples was measured using digital four probe tester in type ST2258C.

Monitoring of current monochromatic pulse responsivity (S_{pulse}) was carried out by method of comparing responsivity of the investigated PD with a reference photodiode certified by the respective metrological service of the company. Measurements were performed when illuminating the PD with a radiation flux of a power of not over 1·10^{-3} W; load resistance across the responsive element R_l= 10 kΩ, at the operating voltage of U_{bias} = 120 V and pulse duration τ_i = 500 ns. The responsivity of samples after diffusion of phosphorus was measured by simulating the reflection of radiation from the gold layer on the reverse side.

To control the resistivity of the end crystals of photodiodes, there is a need to find methods of non-destructive measurement of this parameter, since the measurement by the four-probe method requires etching of the oxide film and the surface doped layers of silicon. We proposed a method of determining the resistivity by studying the width of the space charge region (SCR) (W_i), since these parameters are related by equation (1) [15]:

$$W_i = (\frac{2\varepsilon\varepsilon_0\,(\phi_c - U_{bias})}{eN_A})^{\frac{1}{2}} \qquad (1)$$

where ε, ε_0 is dielectric constants for silicon and vacuum, respectively; φ_c is contact potential difference; U_{bias} is bias voltage; N_A is concentration of acceptors in the substrate.

Note that equation (1) is difficult to calculate, so it is better to use empirical formula (2), which correlates well with experimental data [16]:

$$W_i = \frac{\sqrt{\rho U_{bias}}}{3} \times 10^{-4} \qquad (2)$$

According to the formula (2), it is possible to determine the resistivity:

$$\rho = \frac{9W_i^2}{10^{-8}U_{bias}} \qquad (3)$$

The determination of the width of the space charge region was carried out by measuring the dependence of $S_{pulse}(U_{bias})$, in particular, it was investigated at which voltage the responsivity reached a maximum, accordingly, the width of the SCR in this case reached the thickness of the substrate, since the responsivity reaches saturation when the SCR expands to the entire thickness of the substrate.

The maximum value of the width of the space charge region ($W_{i\,max}$) can be determined from the formula (4):

$$W_{i\,max} = x - (x_{n+-p} + x_{p+-p}) \qquad (4)$$

The dark currents (I_d) of photodiodes were measured using a hardware-software complex implemented on the basis of the Arduino platform, an Agilent 34410A digital multimeter and a Siglent SPD3303X programmable power source, which were controlled by a personal computer using software created by the authors in the LabView environment.

III. Result and Discussion of the Research

Samples of PD$_1$ were studied after the operation of driving-in of phosphorus and after the operation of boron diffusion, as well as the degradation of ρ depending on the combination of thermal operations in the PDs manufacturing route. After the operation of thermal oxidation, it is impossible to determine the resistivity by the proposed method, so the degree of degradation of ρ was determined by measuring this parameter by the 4-probe method. It was established that during this operation, the resistivity of silicon decreases by 1-1.5 kΩ·cm. In fact, the reduction of the resistivity occurs due to contamination of the plates with uncontrolled impurities of metals from the quartz ware, carrier gases or the reactor. With a significant degree of contamination of the quartz reactor, this indicator may increase. To avoid an increase in the rate of degradation of the ρ, the reactors should be periodically purged with hydrochloric acid vapors, etched in hydrofluoric acid or use other methods of quartz purification [14, 17, 18].

A. *Degradation of resistivity after the operation of driving-in of phosphorus*

A graph of $S_{pulse}(U_{bias})$ for the control samples of PD$_1$ after driving-in of phosphorus with the same concentration of doped phosphorus and different density of dark currents was obtained (Fig. 3)(dark currents were measured at a reverse bias voltage of -120 V). Curve 1 characterizes the control sample №1 with density of dark currents J_d=490-520 nA/cm^2. Curve 2 characterizes sample №2 with J_d=600-650 nA/cm^2. Curve 3 characterizes sample №3 with J_d=1040-1300 nA/cm^2. Curve 4 characterizes a sample №4 with J_d=2080-2340 nA/cm^2. We also obtained the dependence of responsivity on the voltage of sample №3 after repeating the phosphorus driving-in operation (sample №3').

It can be seen from the fig. 3 that the dependence curves reach saturation at different bias voltages, this indicates the expansion of the SCR over the entire thickness of the substrate at different voltages, and accordingly, these samples had different resistivity. Curve 1 reaches saturation at U_{bias}=90-95 V, which according to (2) corresponds to ρ≈17.8-18.7 kΩ·cm. Curve 2 reaches saturation at U_{bias}=100-105 V, which corresponds to ρ≈16-16.9 kΩ·cm. Curve 3 reaches saturation at U_{bias}=160-170 V, which corresponds to ρ≈9.9-10.5 kΩ·cm. As for sample №3', after repeated driving-in of phosphorus, the maximum value of responsivity was reached at U_{bias}=200-210 V, which indicated a decrease in resistivity. In this case, the resistivity of the sample reached ρ≈8-8.4 kΩ·cm. From this it can be concluded that the operation of driving-in of phosphorus reduces the specific resistance of the sample by 2 kΩ·cm. Note, that after the repeated thermal operation, the density of the dark current of this sample reached 1560-1820 nA/cm^2. As for sample №4, at U_{bias}=300 V, the responsivity curve has not yet reached saturation (there is no possibility of using a higher bias voltage), accordingly, the proposed method of determining the resistivity is impossible. When measuring the ρ by the 4-probe method, it was established that the specific resistance of this sample reaches 4.5-5.5 kΩ·cm.

It can be seen from the description that the degree of degradation of the resistivity of silicon can also be estimated by the value of the dark current, since samples with a higher resistivity had lower dark currents. This can be explained by the fact that the value of the volumetric generation component of the PDs dark current (I_d^G) is directly proportional to the life time of minor charge carriers (5) [19], the value of which decreases as well as resistivity due to high-temperature treatments.

$$\mathrm{I}_d^G = \mathrm{e}\frac{n_i}{2\tau} W_i A_{\mathrm{RE}} \qquad (5)$$

where e is electron charge; n_i is intrinsic concentration of charge carriers in the substrate, A_{RE} is the area of the RE.

Note that the W_i and, accordingly, the resistivity can also be determined by measuring the *I-V* charateritic, since the value of the dark current reaches saturation, as in the case of responsivity, when the SCR is expanded over the entire thickness of the substrate. To confirm this fact, *IV* characteristics of sample №1-4 was obtained (Fig. 4).

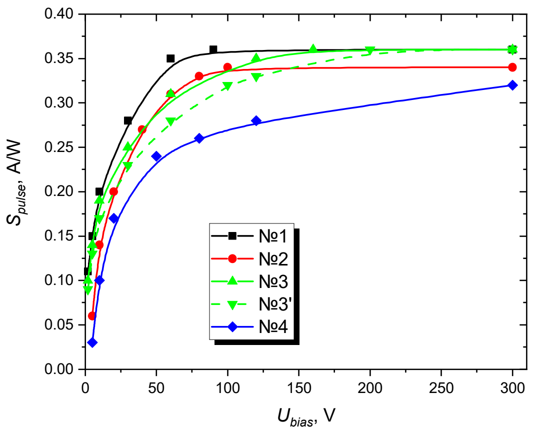


Fig. 3. A graph of $S_{pulse}(U_{bias})$ of the PDs$_1$ after driving-in of phosphorus.

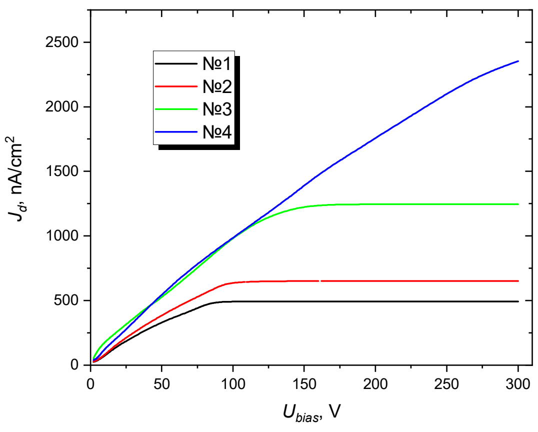


Fig. 4. *I-V* characteristics of the PDs$_1$ after driving-in of phosphorus.

It can be seen from Fig. 4 that the level of dark current reaches saturation at approximately the same voltage as S_{pulse} (Fig. 3). And in the case of sample №4, the *I-V* characterictic at a voltage of 300 V has not yet reached its maximum value, as in the case of its responsivity (Fig. 3 (curve 4)). The W_i can be determined from the equation (5). Different values of dark currents are caused by varying degrees of contamination of the material with uncontrolled impurities, since there are many factors in the PDs manufacturing technology that can contribute to this, in particular, possible contamination from the quartz vessel or reactor, from carrier gases, etc.

B. Degradation of resistivity after the operation of diffussion of boron

The degradation of silicon resistivity after boron diffusion was also investigated. To determine the value of ρ, the dependences of $S_{pulse}(U_{bias})$ of the final crystals of PDs from batches of samples №1 and №4 were obtained by the proposed non-destructive method (Fig. 5). Let's call the tested samples №1^B and № 4^B.

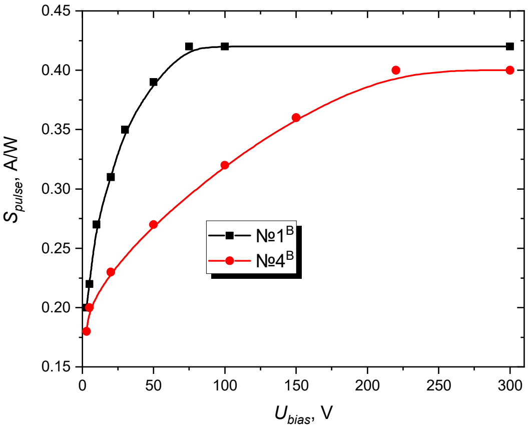


Fig. 5. A graph of $S_{pulse}(U_{bias})$ of the PDs$_1$ after diffussion of boron.

From Fig. 5, it can be seen that the curve for sample №1^B reaches saturation at U_{bias}=75-80 V, which corresponds to ρ≈21-22.5 kΩ·cm in contrast to ρ≈17.8-18.7 kΩ·cm of sample №1 after driving-in of phosphorus. Curve №4^B reaches saturation at U_{bias}=210-220 V, which corresponds to ρ≈7.7-8 kΩ·cm, in contrast to ρ≈4.5-5.5 kΩ·cm of sample №4 after driving-in of phosphorus. From the above, it can be concluded about some recovery of the resistivity of the silicon due to gettering of the reverse side of the plates with boron. In the investigated case, it was possible to restore the resistivity of silicon by 2-4 kΩ·cm relative to the value after phosphorus diffusion. The recovery rate of the resistivity depends on the duration of the gettering operation and the concentration of doped boron [20].

Also, from Fig. 5 it can be seen that the overall level of responsivity has increased in both cases, this is also due to the recovery of the lifetime of the minor charge carriers due to gettering. We also note that to measure the S_{pulse} of the samples after phosphorus driving-in, an anti-reflective SiO_2 was etched, in contrast to the responsivity measurement after boron diffusion, where oxide etching was not carried out, accordingly, a decrease in the reflection coefficient from the surface of the samples also contributes to an increase in responsivity, which can be seen from equation of responsivity of a real photodiode (6) [11]:

$$S_\lambda = (1-R)TQ\alpha_{p-n}\frac{\lambda}{1.24} \tag{6}$$

where R is the reflection coefficient of the crystal surface; T is the transmission coefficient of the input window or optical filter; Q is the quantum output of the internal photoeffect, α_{p-n} is charge carrier collection coefficient.

C. Degradation of resistivity when using mesa-technology and liquid phase diffusion

When we manufacture photodiodes using mesa technology (PD$_2$) and using one-stage diffusion using PCl_3 (PD$_3$), it is possible to reduce the number of thermal operations in the technological routes, in particular, to exclude operations with T=1423 K. To establish the degree of degradation of the specific resistance when using the specified technologies, we obtained graphs of $S_{pulse}(U_{bias})$ (Fig. 6).

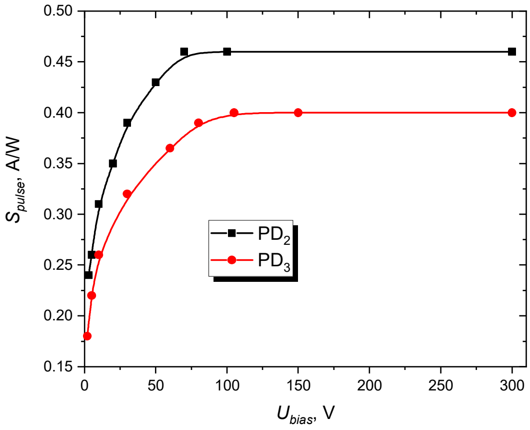


Fig. 6. A graph of $S_{pulse}(U_{bias})$ of the PD$_2$ and PD$_3$.

From Fig. 6 it can be seen that the responsivity curve of PD$_2$ reaches saturation at a voltage of 70-75 V, which corresponds to ρ≈22.5-24 kΩ·cm, that is, when using mesa-technology, minimal degradation of the resistivity of silicon is possible. As for PD$_3$, in this case, the degree of degradation of resistivity is somewhat worse than samples of type PD$_1$ and PD$_2$. The resistivity of completed crystals of the PD$_3$ type reached 16-17 kΩ·cm. This is caused by significant defect formation on the surface of the substrates during diffusion from the liquid phase of phosphorus chloride. During our study of the structural perfection of the surface of doped samples of the PD$_2$-type after selective etching, dislocation grids with a high density were revealed (Fig. 7).

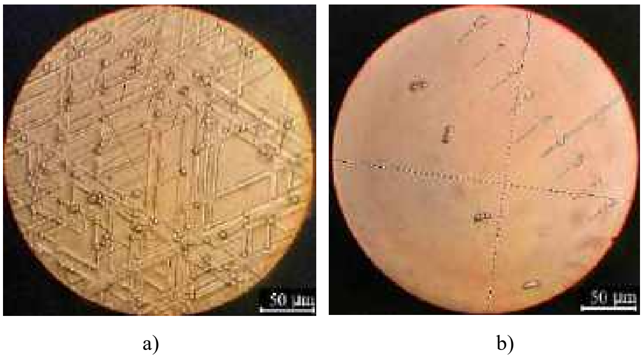


Fig. 7. Image of dislocations on the surface of plates after phosphorus diffusion using PCl_3 (a) and planar sources (b).

The main cause of the formation of a high density of structural defects in the case of liquid-phase diffusion is the location of a significant number of phosphorus atoms in the internodes of the silicon crystal lattice. These atoms are electrically inactive. Accordingly, these impurities introduce significant mechanical stresses caused by the difference in the sizes of phosphorus and silicon atoms, which leads to formation of dislocations [19] (Fig. 7a). When using phosphorus diffusion from planar sources, it is possible to obtain a significantly lower density of structural defects (Fig. 7b) compared to liquid phase diffusion [14, 21].

It can also be seen from Fig. 6 that the responsivity of PD_3 is significantly lower than PD_1 and PD_2. This is caused by a high number of generation-recombination centers - dislocations formed during the diffusion of phosphorus, which reduce the lifetime of minor charge carriers. In samples of the PD_2-type, the value of responsivity is the highest among all considered cases of the technology, respectively, the degree of degradation of the life time of minor charge carriers and the resistivyty is the lowest. Accordingly, the use of mesa-technology allows to minimize the degradation of the electrophysical characteristics of silicon by reducing the number of thermal operations. Note that the use of slow cooling after heat treatments also avoids deterioration of material characteristics.

Conclusions

The degradation of silicon resistivity in the process of manufacturing silicon four-element *p-i-n* photodiodes with a guard ring was investigated and the following conclusions were drawn. A non-destructive method of controlling the resistivity of finished crystals by determining the reverse bias voltage, at which the space charge region acquires the width of the substrate, is proposed. This voltage can be determined from graphs of responivity versus voltage or current-voltage characteristics. the proposed resistivity measurement method correlates well with the 4-probe method. The reduction of the specific resistance occurs due to contamination of the plates with uncontrolled impurities of metals from the quartz ware, carrier gases or the reactor. Operations of oxidation and driving-in of phosphorus will reduce the resistivity of silicon by 1-1.5 kΩ·cm and 2 kΩ·cm, respectively. During the operation of boron gettering of the reverse side of substrate, some restoration of the resistivity occurs. In the investigated case, it was possible to restore the resistivity of silicon by 2-4 kΩ·cm relative to the value after phosphorus diffusion. The degradation of resistiwity acquires a significant character when using diffusion of phosphorus from phosphorus chloride due to the formation of a high density of dislocations. The use of mesa-technology allows to minimize the degradation of the electrophysical characteristics of silicon by reducing the number of thermal operations.

References

[1] N. Margalit, C. Xiang, S. M. Bowers, A. Bjorlin, R. Blum, J. E. Bowers, (2021). Perspective on the future of silicon photonics and electronics. Applied Physics Letters, 118(22), 220501. https://doi.org/10.1063/5.0050117

[2] P. Zhang, J. Duan, G. Chen, J. Li, W. Wang, (2018). Production of polycrystalline silicon from silane pyrolysis: A review of fines formation. Solar Energy, 175, 44-53. https://doi.org/10.1016/j.solener.2017.12.031

[3] H. Xu, (2015). Characterization of n-type mono-crystalline silicon ingots produced by continuous Czochralski (Cz) Technology. Energy Procedia, 77, 658-664. https://doi.org/10.1016/j.egypro.2015.07.095

[4] Y. B. Vasilév, N. A. Verezub, M. V. Mezhennyi, V. S. Prosolovich, A. I. Prostomolotov, V. Y. Reznik, (2013). Features of defect formation under the thermal treatment of dislocation-free single-crystal large-diameter silicon wafers with the specified distribution of oxygen-containing gettering centers in the bulk. Russian microelectronics, 42, 467-476. https://doi.org/10.1134/S1063739713080155

[5] C. Kranert, M. Trempa, C. Reimann, J. Friedrich, (2021). Metal contamination of silicon from the furnace atmosphere after crystallization. Journal of Crystal Growth, 559, 126026. https://doi.org/10.1016/j.jcrysgro.2021.126026

[6] Y.V. Meteleva-Fischer, A.J. Böttger, W.G. Sloof, et al. Gas-Phase Interactions as Sources of Contamination in Solar Silicon. Metallurgical and Materials Transactions E 1, 174–179 (2014). https://doi.org/10.1007/s40553-014-0017-6

[7] G. P. Gaidar (2018). The influence of different heat treatment regimes on the Hall parameters and lifetime of charge carriers of transmutationally doped silicon crystals. Journal of Physical Research, (22, No. 4), 4601-1. https://doi.org/10.30970/jps.22.4601 [in Ukrainian].

[8] W. S. Yoo, T. Fukada, I. Yokoyama, K. Kang, N. Takahashi, (2002). Thermal behavior of large-diameter silicon wafers during high-temperature rapid thermal processing in single wafer furnace. Japanese journal of applied physics, 41(7R), 4442. https://doi.org/10.1143/JJAP.41.4442

[9] M. S. Kukurudziak, E. V. Maistruk, (2023). High-responsivity silicon pin mesa-photodiode. Semiconductor Science and Technology, 38, 085007. https://doi.org/10.1088/1361-6641/acdf14

[10] K.O. Boltar, I.V. Chinareva, A.A. Lopukhin, N.I. Yakovleva, (2013) Matrix planar and mesa-structures based on heteroepitaxial InGaAs layers. Applied Physics. 5, 10-15.

[11] A.V. Fedorenko, Spectral responsivity of diffused Ge p-i-n photodiodes. Tekhnologiya i Konstruirovanie v Elektronnoi Apparature. (2020). 3–4, 17 - 23.

http://dx.doi.org/10.15222/ TKEA2020.3-4.17 [in Ukrainian].

[12] M. S. Kukurudziak, (2022). Influence of Surface Resistance of Silicon pin Photodiodes n+-Layer on their Electrical Parameters. Physics and Chemistry of Solid State, 23(4), 756-763. https://doi.org/10.15330/pcss.23.4.756-763

[13] M. S. Kukurudziak, (2023). Problems of chemical and dynamic polishing in the technology of silicon pin photodiodes. Chemistry, Physics & Technology of Surface/Khimiya, Fizyka ta Tekhnologiya Poverhni, 14(1). https://doi.org/10.15407/hftp14.01.042

[14] M. S. Kukurudziak, (2022). Diffusion of phosphorus in technology for manufacturing silicon pin photodiodes. Semiconductor Physics, Quantum Electronics & Optoelectronics, 25(4), 385-393. https://doi.org/10.15407/spqeo25.04.385

[15] N. M. Tugov, B. A. Glebov, N. A. Charykov, Poluprovodnikovyye pribory: Uchebnik dlya vuzov [Semiconductor devices: Textbook for universities], Ed. V. A. Labuntsov, Energoatomizdat, Moscow, 1990. 576 p. [in Russian]

[16] V.A. Bruk, V.V. Garshenin, A.I. Kurnosov Production of semiconductor devices: Textbook. Ed. 3rd, revision and supplement. M.: Vysshaya Shkola, 1973.-264 p. [in Russian]

[17] X. Pan, S. Li, Y. Li, P. Guo, X. Zhao, Y. Cai, (2022). Resource, characteristic, purification and application of quartz: a review. Minerals Engineering, 183, 107600. https://doi.org/10.1016/j.mineng.2022.107600

[18] W. Song, X. Jiang, C. Chen, B. Ban, S. Wan, J. Chen, (2023). Purification of Quartz Via Low-Temperature Microwave Chlorinated Calcination Combined with Acid Leaching and its Mechanism. Silicon, 15(2), 971-981. https://doi.org/10.1007/s12633-022-01749-w

[19] K. V. Ravi, Imperfections and impurities in semiconductor silicon (N.Y.: Wiley, 1981).

[20] M. S. Kukurudziak and E. V. Maistruk, "Features of Diffusion Doping and Boron Gettering of Silicon p-i-n Photodiodes," 2022 IEEE 3rd KhPI Week on Advanced Technology (KhPIWeek), 2022, pp. 1-6, doi: https://doi.org/10.1109/KhPIWeek57572.2022.9916420.

[21] M.S. Kukurudziak (2022). Formation of Dislocations During Phosphorus Doping in the Technology of Silicon p-i-n Photodiodes and their Influence on Dark Currents. Journal of nano- and electronic physics. Vol. 14 No 4, 04015(6cc). https://doi.org/10.21272/jnep.14(4).04015

Defect Formation on the Surface of Silicon Substrates after Various Technological Operations

Mykola S. Kukurudziak
Engineering production complex 1
Rhythm Optoelectronics Shareholding Company
Chernivtsi, Ukraine
mykola.kukurudzyak@gmail.com

Volodymyr M. Lipka
Rhythm Optoelectronics Shareholding Company
Chernivtsi, Ukraine
volodymyrlipka9@gmail.com

Olha P. Andreeva
Engineering production complex 1
Rhythm Optoelectronics Shareholding Company
Chernivtsi, Ukraine
olgaandryeyewa@gmail.com

***Abstract* — The article describes the defect formation on the surface of silicon substrates during various technological operations in the manufacture of electronics elements. It is established that when the substrates are separated into elements/crystals by cutting with a diamond disk, dislocation grids are formed along the cutting line. It is also shown that the use of liquid-phase diffusion of phosphorus contributes to more intense defect formation than diffusion from flat sources. It has been shown that silicon with crystallographic orientation [100] is more prone to defect formation during technological operations than silicon with orientation [111]. It has also been found that defects in the form of craters can form on the surface of plates during thermal gold sputtering as a result of local melting of silicon when gold droplets hit substrates with a temperature higher than the melting point of silicon. Some migration of dislocations to the periphery of the crystal in the presence of inversion layers after isothermal annealing was found.**

Keywords — silicon, crystallographic defect, dislocation, technological operation.

I. Introduction

The production of electronic devices and integrated circuits based on silicon involves a number of complex chemical, physical and thermal processes. In brief, the manufacturing of semiconductor devices is as follows. Purification of industrial silicon, which is a product of the reduction of pure quartz sand. Such purified polycrystalline silicon is further used to grow doped single crystals. Single-crystal cylindrical ingots are cut into round, thin plates, which are mechanically and chemically polished to produce a smooth, flat surface without mechanical damage. Next, various alloying elements of both types of conductivity are locally diffused into the single-crystal plates. The localization of diffusion in different areas of the crystal surface is ensured by applying a protective silicon oxide film obtained by thermal oxidation and subsequent photolithography in the corresponding areas of the surface. The electrical connection of individual areas of the devices to each other is carried out using thin metal films obtained by vacuum sputtering. Usually, a large number of individual devices or circuits are simultaneously formed on a single-crystal silicon wafer, which are separated into crystals by scribing or cutting with a diamond-edged disk, and then by enclosure operations [1].

It is worth noting that the main problem of the described technological process is the production of single-crystal ingots and wafers with a minimum number of crystallographic defects, since their presence negatively affects the electrical properties and parameters of manufactured devices [2, 3]. It should be noted that the use of defect-free material does not ensure the absence of defects in the final products, because during mechanical and thermal processes in the manufacture of electronics elements, violations of the crystallographic lattice of silicon are observed. Investigating the mechanisms of defect formation on the surface of silicon substrates after various technological operations and establishing methods to avoid them is an important and urgent scientific and technical task.

Methods for producing defect-free single-crystal silicon ingots and dislocation-free wafers, as well as the problems of defect formation during ingot growth, are known and well-studied [4-7]. In particular, [5] indicates that when large-diameter Si crystals are drawn from the melt, small changes in the drawing rate and temperature gradient can lead to the formation of voids that cause damaging pits on the surface of a polished wafer. Also, [6] investigated the effect of growth impurities (oxygen and carbon) on the magnetic, micromechanical, and structural properties of silicon single crystals grown by the Czochralski method (Cz-Si) after their heat treatment in the temperature range of 700 - 1100 ˚C. In [7], the authors analyzed the influence of the pickle structure on the further growth of the ingot and emphasized the importance of using defect-free pickles. However, the mechanisms of defect formation on the surface of silicon substrates during various technological operations in the manufacture of electronic devices have not been sufficiently investigated. Accordingly, the study of the causes of defect formation on the surface of silicon wafers in the technological process and methods of their avoidance is the aim of this work.

II. Experimental Details

The research was carried out on the basis of the technological route for the manufacture of silicon *p-i-n* photodiodes (PDs). The PDs technological route consisted of a complex of four thermal operations and three photolithographies: after machining of the silicon wafers, the chemical-dynamic polishing (CDP) was performed according to the technological modes given in [8]; than semiconductor substrates were oxidized according to the principle of dry-wet-dry oxidation; photolithography was carried out to create windows for phosphorus diffusion; diffusion of phosphorus (predeposition) to the front side to create n^+-type responsive elements (using planar sources of phosphorous); driving-in of phosphorus to redistribute the alloying impurity and increase the depth of the n^+-p-junction at a temperature; diffusion of boron to the reverse side of the substrate to create a p^+-type ohmic contac; photolithography for creating contact windows; sputtering of Cr-Au on the front and back sides; separation of pastes into crystals by scraping or cutting with a disk with an external diamond edge.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

Given that we described defect formation on the surface of substrates during thermal oxidation in [9], in this work we will pay special attention to the operations of phosphorus diffusion, CDP, cutting substrates into crystals, and metallization. It should be noted that to compare the effect of diffusion methods on the crystallographic structure of wafers, we also performed phosphorus diffusion from a PCl_3 liquid-phase diffuser according to the technological modes described in [10]. The surface concentration of phosphorus in all versions of the technology reached N_{P0}=4.1-4.15·10^{20}cm^{-2} (R_s≈2.7-2.8 Ω/□)

For chemical-dynamic polishing, an etchant with the following composition was used:

$$HNO_3 : HF : CH_3COOH = 9 : 2 : 4.$$

Dislocation-free *p*-type single crystal FZ-Si with [111] and [100] orientation, resistivity ρ≈12-24 kΩ·cm was used.

To investigate the defective structure of the substrates with [111] orientation, chemical treatment was performed in selective Sirtle`s etchant [11] with the following composition:

$$HF – 100\ cm^3,\ CrO_3 – 50\ g,\ H_2O – 120\ cm^3$$

To investigate the defective structure of the substrates with [100] orientation, chemical treatment was performed in selective Secko's etchant [12] with the following composition:

$$4.4\%\ K_2Cr_2O_7 : HF = 1 : 2$$

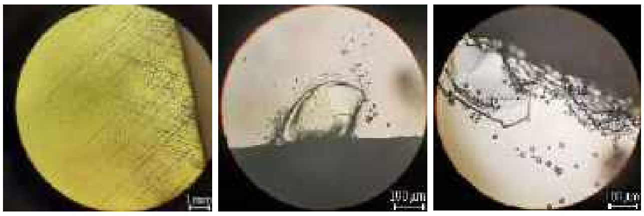
Then the surface was examined in microscopes of different magnifications.

III. Result and Discussion of the Research

A. Defect formation when cutting substrates into elements/crystals

Depending on the technological needs, silicon wafers can be separated into elements before and after thermal operations. When cutting wafers with a diamond-edged blade before heat treatment, dislocations were observed along the cutting lines (Fig. 1). The formation of dislocations was provoked by mechanical stresses that occur during during disk cutting. During the selective etching of of the samples that were separated after thermal operations, no dislocation clusters were detected. It can be argued that the formation of structural defects in the first case was caused by high temperatures and diffusion of impurities, when areas with increased mechanical stresses became places of localization of dislocations formed as a result of stress relaxation at high temperatures [13]. Dislocation clusters were also found in places of chips along the cutting line of silicon wafers, which indicates the presence of local mechanical stresses in the chip areas. The formation of chips is possible when using an unsharpened disk. In order to avoid the influence of defects on the photovoltaic parameters of products, it is worth applying the cutting operation after thermal operations. The distance from the active elements of the device to the cutting lines is also important. By increasing this distance, it is possible to minimize the degradation of the parameters of electronics elements caused by the described defect formation.

It should be noted that when using the scraper method of separating wafers along cutting and chipping lines, significant chipping of silicon can occur, which leads to the formation of a high density of dislocations and other structural defects (Fig. 1).



a) b) c)

Fig. 1. Defect formation on the surface of the plates during cutting with a diamond-edged blade (a) and scraping (b, c).

As can be seen in Fig. 1, the use of scribing leads to significant defect formation, and therefore the use of this method is impractical. However, it should be noted that in recent decades, the method of laser controlling thermal scraping has become widespread in recent decades [14]. The process is based on the heating carried out by a laser beam along the cutting line on the plate with subsequent cooling of the heated area with a refrigerant. Laser irradiation of the surface of the material leads to the emergence of significant compressive stresses in its outer layers of significant compression stresses, which do not directly lead to fracture. But it is thanks to the refrigerant that a sharp local cooling and the resulting temperature gradient leads to the appearance in the tensile stresses in the surface layer, which create conditions for the formation and progression of cracking. However, this cutting method has a number of requirements for to ensure high quality processing, which complicate this process: the need to mount the plate on a satellite film, the inadmissibility of films on the cutting tracks photoresist or other organic substances on the cutting tracks, the need to protect the plate with a polymeric coating during gluing to the film satellite to avoid condensation on the evaporated material on the surface, etc. [15]. This method is not suitable for high-volume silicon production due to its significant labor intensity. For this purpose, cutting with a diamond blade with an outer cutting edge, but it is necessary to control the degree of degradation of the cutting blade and take into account the distance from the cutting line to the active elements of the products.

B. Defect formation when chemical dynamic polishing

The CDP operation is designed to remove the disturbed layer of the plate formed during mechanical processing and chemical-mechanical etching. It is worth noting that during grinding and polishing operations, surface defects or inclusions of another phase can be "hidden". If the thickness of the chemical dynamic polishing is insufficient, hidden defects can be detected, which leads to uneven etching and polishing and, as a result, a violation of the plane parallelism of the plates. During the selective etching of plates with point inclusions of another phase, it was seen that these defects are areas of high dislocation density (Fig. 2). Accordingly, to prevent disruption of the surface structure of the plates, it is necessary to perform etching to a depth that completely removes the disturbed layer. In the mass production of photodetectors, we remove the surface layer of wafers with a thickness of up to 25 microns before thermal operations.

It should be noted that changing the concentration of the components of the CDP etchant significantly changes its properties – the etchant can become selective. One of the well-known selective etching agents is Desch's etchant [7]. Its composition is similar to the polishing solution:

$$HNO_3 : HF : CH_3COOH = 3 : 1 : (8\text{-}12)$$

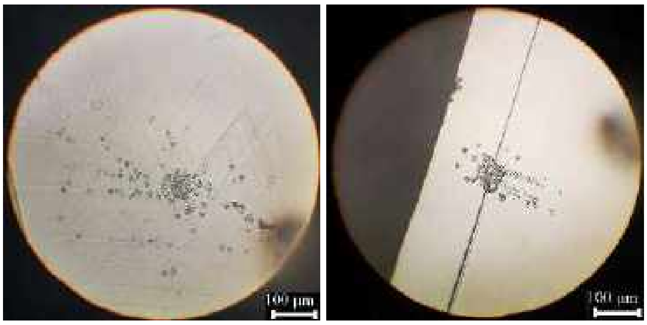


Fig. 2. Image of dislocation accumulation on the surface of silicon wafers in the places of inclusion of another phase after CDP.

When polished with such an etchant, the plates become matte (Fig. 3a), or with a lower degree of selectivity of the etchant, the surface of the substrates can be covered with the so-called "starry sky" - dots that appear when examined in a dark-field microscope. Examination of the substrates in high-magnification microscopes revealed triangular etching pits on their surface, which are characteristic of silicon with a crystallographic orientation [111] (Fig. 3b).

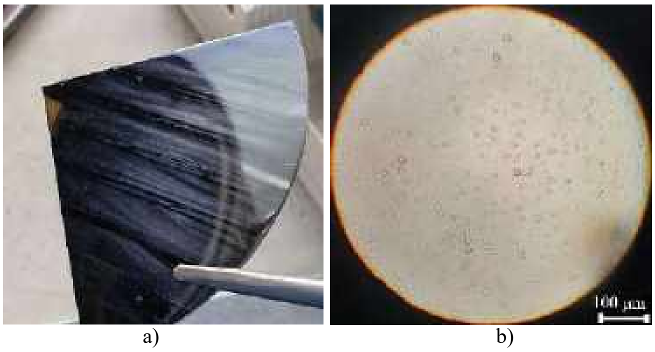


Fig. 3. Images of the matte surface (a) and triangular etching pits (b) formed as a result of the etchant's selective properties.

The etching pits shown in Fig. 3b are not edge dislocations, since the formation of etch pits is observed even when polishing dislocation-free silicon. In the case of polishing substrates with an etchant with selective properties, the presence of crystallographic disorders is not a prerequisite for the appearance of etch pits, but the density of pits increases in places where structural defects or mechanical stresses are localized. To avoid the described defect formation, it is necessary to strictly observe the proportions when preparing the pickling agent and to conduct incoming inspection of chemical reagents.

C. Defect formation on the surface of the plates during the diffusion of impurities

The mechanisms of defect formation, in particular dislocations on the surface of silicon wafers during impurity diffusion, are well studied and investigated [1,3]. The main factors that contribute to the formation of dislocations in this case are thermomechanical stresses that occur in the wafers during rapid heating or cooling and the actual introduction of atoms with larger or smaller atomic radii into the crystallographic lattice of the base material, which causes mechanical stresses. However, some new features of dislocation formation during phosphorus diffusion have been revealed. In particular, it was found that the density of dislocations when using diffusion from solid-state planar sources is much lower than when using diffusion with a liquid diffuser PCl_3 at the same surface concentration of the impurity (Fig. 4).

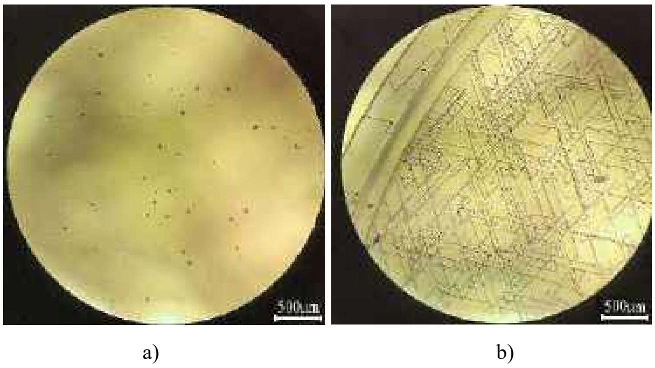


Fig. 4. Image of the defective structure on the surface of silicon wafers after phosphorus diffusion using planar sources (a) and PCl_3 (b).

The density of dislocations on the plate surface shown in Fig. 4a reaches N_{dis}≈$2{\cdot}10^3$ – $3{\cdot}10^3$ cm^{-2}, and in Fig. 4b N_{dis}≈$3{\cdot}10^8$ – $4{\cdot}10^8$ cm^{-2}. Such a difference in dislocation density is caused by the fact that in the process of diffusion from the liquid phase, a significant amount of impurities enters the interstices of the crystal lattice, and they are not electrically active, unlike diffusion from planar phosphorus sources. Accordingly, these impurities introduce significant mechanical stresses caused by the difference in the size of of phosphorus and silicon atoms, which leads to dislocation formation.

We have also seen that with a significant duration of selective etching (more than 15 minutes) of silicon wafers with crystallographic orientation [111], the etching pits corresponding to edge dislocations are modified. The edges of the classical quasi-pyramidal pits with a triangular base are etched to form cubic pits with a mutual placement of the edges of 120 degrees (Fig. 5). In the study of silicon wafers with orientation [100] doped with phosphorus after selective etching, it was found that the material with this crystallographic orientation is much more prone to the formation of dislocations (Fig. 6).

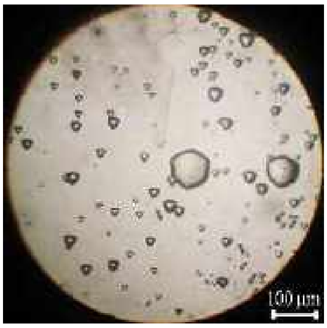


Fig. 5. Surface of a silicon wafer after selective etching for 15 minutes.

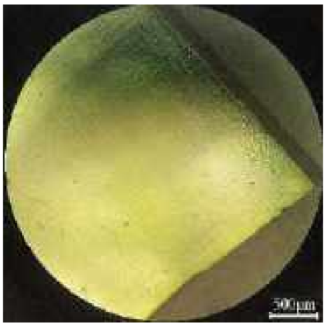


Fig. 6. Surface of a silicon wafer of orientation [100] doped with phosphorus after selective etching.

As can be seen from Fig. 6, the surface of the doped area is completely covered with a dense dislocation network, which makes the calculation of the defect density difficult, but it can be visually assessed that, compared to Fig. 4, the dislocation density in this case is much higher (the impurity concentration is the same). Presumably, the increased susceptibility of such a material to defect formation is caused by the reduced concentration of atoms per unit area compared to the orientation [111].

D. *Defect formation during thermal spraying of gold*

In the mass production of silicon products, some degradation of parameters was observed after the gold sputtering stage. An assumption was made about the formation of defects on the silicon surface. To assess the structural perfection of the substrate surface, they were examined after etching in a selective etchant. On the surface of the rejected crystals, we found complexes of structural defects in the form of craters, which are clusters of randomly placed dislocations and point defects (Fig. 7). Complexes of structural defects of this type were found for the first time on the surface of silicon wafers.

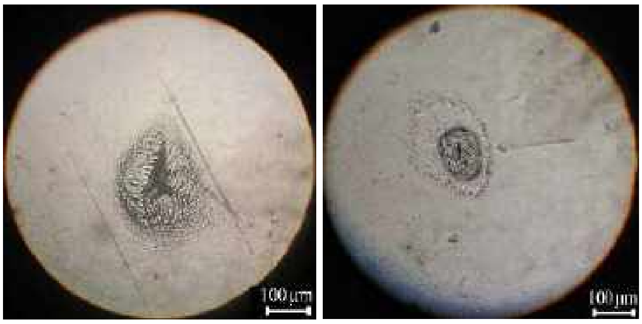


Fig. 7. Defects caused by localized melting of silicon.

Such defects are formed as a result of local melting of silicon when "drops" of gold boiling in the evaporator with a temperature higher than the melting point of silicon hit it. In the places of their localization, a thickened, sometimes sharp-edged layer of gold is observed, which requires longer etching during photolithography, leading to etching and, as a result, to a change in the specified shape and size of the contact pads. In addition, such defects can damage photo templates by forming scratches. As a result of the research, it was found that wire sputtering is accompanied by more intense "clogging" of substrates with gold droplets than when using kings. The mechanism of this phenomenon requires additional research. The likelihood of the described defects can be reduced by sputtering from closed evaporators or by increasing the sputtering time per flap during gold melting. However, it should be borne in mind that these methods increase the consumption of precious metal.

E. *Dislocation migration on the surface of silicon wafers with inversion layers*

When studying silicon wafers with surface inversion layers at the semiconductor-oxide interface, some dislocation dynamics was observed. It was manifested in the migration of these structural defects to the periphery of the crystal and their localization with high density in areas outside the sensitive elements of photodetectors (Fig. 8). It should be noted that the wafers processed according to the same technological route but without the existing inversion layers had dislocations evenly distributed on the surface. This phenomenon contributed to a significant reduction in the density of defects in the active areas of the products.

It is known from sources that the main driving force for the movement of dislocations on the surface of plates is the stresses introduced by deformation [16, 17], but in this case there is no mechanical impact on the substrates. The described movement of dislocations on the surface of silicon substrates during isothermal annealing caused by the formation of inversion layers was detected for the first time.

The mechanism of the observed dynamics of dislocations on the surface of plates with inversion is not yet known to us, so this phenomenon requires additional research.

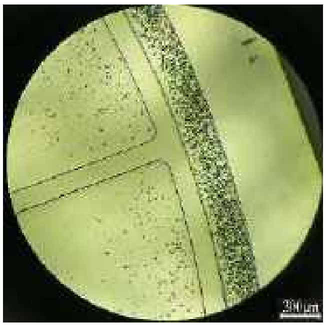


Fig. 8. Localization of dislocations on the periphery of crystals due to migration.

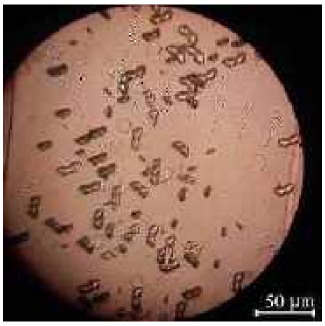


Fig. 9. Images of edge dislocations and hexagonal Frank dislocation loops.

When studying the morphology of the areas localization of defects after migration in microscopes with high magnification, the formation of of hexagonal and circular defects (Fig. 9), which are partial edge dislocation Frank loops [1, 18]. It should also be noted that the formation of hexagonal edge dislocation Frank loops was not detected after oxidation or diffusion, but their generation was observed after the formation of inversion layers and the described dislocation migration, which was also established for the first time.

F. *Some other factors of defect formation on the plate surface*

When the silicon oxide was removed from the wafers, chaotic complexes of structural defects were found on their periphery, which were the so-called furrow-like formations (Fig. 10a-d) and etching pits (Fig. 10e). The sources do not describe the mechanisms of formation of the described disturbed surface, but when investigating the causes of these defects, it was found that the places of their localization are unused areas of the substrates, where high-quality photolithography is not required. Accordingly, in these areas, the silicon oxide becomes porous due to etching (Fig. 10f). During subsequent chemical treatments, the unprotected areas are etched to form a furrow-like structure, which is clearly visible during oxide film removal and selective etching.

Also, spiral-shaped structural defects were found on the surface of the silicon substrates, which extend deeply into the thickness of the wafers (Fig. 11). These defects can be confused with swirl defects, but the latter are clusters of point defects with a spiral distribution [1, 19], which is not typical of the detected defects. It is worth noting that a review of the scientific literature on defect formation in silicon did not reveal any descriptions of such defects. A possible reason for the appearance of these violations is the point mechanical impact on the plate during machining, another reason may be the inclusion of another phase during the growth of silicon ingots, which provokes the formation of concentric structural defects.

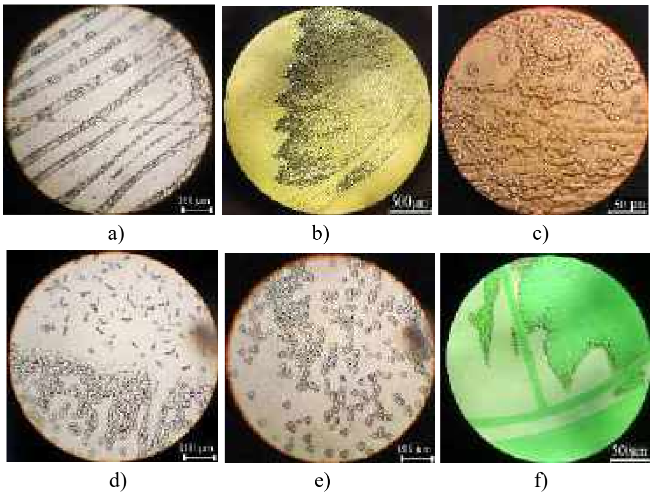


Fig. 10. Complexes of structural defects in the form of furrow-like formations (a-d) and etching pits (e) formed in the areas of porous silicon oxide.

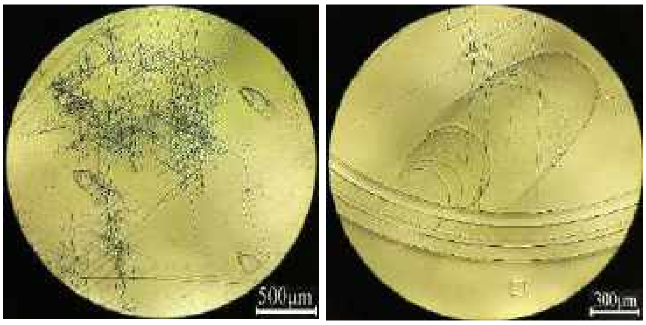


Fig. 11. Complexes of spiral-shaped structural defects.

The described types of structural defects adversely affect the parameters of the manufactured products, and therefore, in the manufacture of silicon electronics, it is necessary to use technological modes that minimize defect formation.

Conclusions

The defect formation on the surface of silicon substrates during various technological operations in the manufacture of electronics elements has been studied and a number of conclusions have been drawn. The use of liquid-phase diffusion of phosphorus using PCl_3 contributes to more intense defect formation than diffusion from flat sources. Silicon with crystallographic orientation [100] is more prone to defect formation during technological operations than silicon with orientation [111]. This is probably due to the reduced concentration of atoms per unit area compared to the orientation [111]. Cratered defects can form on the surface of the plates during thermal gold sputtering as a result of local melting of silicon. Some migration of dislocations to the periphery of the crystal in the presence of inversion layers after isothermal annealing was detected. The formation of a broken or disrupted SiO_2 layer may result in the formation of furrow-like defects on the silicon surface when the oxide film is removed as a result of some etching during chemical treatments. Spiral-shaped structural defects were found on the surface of the silicon substrates, which extend deeply into the thickness of the wafers. A possible reason for the appearance of these violations is the point mechanical impact on the plate during machining.

References

[1] K. V. Ravi Imperfections and impurities in semiconductor silicon. N.Y.: Wiley, 1981 (in Russian).

[2] B.Son, Y. Lin, K. H. Lee, Q. Chen, C. S. Tan, (2020). "Dark current analysis of germanium-on-insulator vertical pin photodetectors with varying threading dislocation density," Journal of Applied Physics, 127(20), 203105. https://doi.org/10.1063/5.0005112

[3] M.S. Kukurudziak, "Formation of Dislocations During Phosphorus Doping in the Technology of Silicon p-i-n Photodiodes and their Influence on Dark Currents," Journal of nano- and electronic physics. Vol. 14 No 4, 04015(6cc) (2022). DOI: https://doi.org/10.21272/jnep.14(4).04015

[4] Arnberg, L., Di Sabatino, M., & Ovrelid, E. J. (2012). State-of-the-art growth of silicon for PV applications. Journal of crystal growth, 360, 56-60. https://doi.org/10.1016/j.jcrysgro.2012.03.024

[5] Kamiyama, E., Vanhellemont, J., Sueoka, K., Araki, K., & Izunome, K. (2013). Thermal stress induced void formation during 450 mm defect free silicon crystal growth and implications for wafer inspection. Applied Physics Letters, 102(8). https://doi.org/10.1063/1.4793662

[6] Pavlovskyy Y., Berbets O., & Lytovchenko P. (2021). Influence of growth impurities on thermal defect formation in monocrystalline silicon. Physics and Chemistry of Solid State, 22(3), 437-443. https://doi.org/10.15330/pcss.22.3.437-443

[7] Liu, S., Huang, X., Wang, Y., Xia, M., Lei, Q., & Zhou, N. (2022). Suppression of dislocations and twins by inducing asymmetrical grain boundaries for casting high-quality monocrystalline silicon ingot. Vacuum, 206, 111533. https://doi.org/10.1016/j.vacuum.2022.111533

[8] M. S. Kukurudziak. (2023). Problems of chemical-dynamic polishing in the technology of silicon p-i-n photodiodes. Surface Chemistry, Physics and Technology, 14(1), 42. https://doi.org/10.15407/hftp14.01.042 (in Ukrainian)

[9] M. S. Kukurudziak (2023) "Problems of Masking and Anti-Reflective SiO2 in Silicon Technology," East European Journal of Physics, (2), 289-295. https://doi.org/10.26565/2312-4334-2023-2-33

[10] M. S. Kukurudziak, (2022). "Diffusion of phosphorus in technology for manufacturing silicon pin photodiodes," Semiconductor Physics, Quantum Electronics & Optoelectronics, 25(4), 385-393. https://doi.org/10.15407/spqeo25.04.385

[11] E. Sirtl, A. Adler (1961) "Flubsaure als sperifishes system zur atzgrubenentwicklang auf silizium," Z. Metallk, 119, 529–31.

[12] F. A. Abdullin, V. E. Pautkin, (2019). "Application of the Selective Silicon Etching Methods for Estimation of the Wafers Quality in the Micromechanical Sensors," In 2019 International Seminar on Electron Devices Design and Production (SED) (pp. 1-4). IEEE. https://doi.org/10.1109/SED.2019.8798467

[13] H. Wu, S. N. Melkote, (2013). "Effect of crystal defects on mechanical properties relevant to cutting of multicrystalline solar silicon," Materials science in semiconductor processing, 16(6), 1416-1421. https://doi.org/10.1016/j.mssp.2013.05.016

[14] J. Xu, H. Hu, C. Zhuang, G. Ma, J. Han, Y. Lei, (2018). "Controllable laser thermal cleavage of sapphire wafers," Optics and Lasers in Engineering, 102, 26-33. https://doi.org/10.1016/j.optlaseng.2017.10.012

[15] V.I. Ivanov, V.S. Kondratenko, (2018). "Modern methods and equipment for cutting instrument wafers into crystals (review)," Advances in applied physics, 6(2), 174. [in Russian]

[16] H. Chen, V.I. Levitas, L. Xiong, X. Zhang (2021). "Stationary dislocation motion at stresses significantly below the Peierls stress: Example of shuffle screw and 60 dislocations in silicon," Acta Materialia, 206, 116623. https://doi.org/10.1016/j.actamat.2021.116623

[17] N. Zhou, X. Wei, L. Zhou, (2018). "Formation of dislocations in the growth of silicon along different crystallographic directions—A molecular dynamics study," Crystals, 8(9), 346. https://doi.org/10.3390/cryst8090346

[18] I.E.Talanin, D.I. Levinzon, V.I. Talanin "Investigation of the transformation of growth microdefects in silicon after ion implantation," Ukr. J. Phys. 2000. 45(8): 963. (in Ukrainian).

[19] Z. Wang, X. Zhu (2022) "Comprehensive characterization of efficiency limiting defects in the swirl-shaped region of Czochralski silicon," Solar Energy Materials and Solar Cells, 236, 111533. https://doi.org/10.1016/j.solmat.2021.111533

Sensors Sensitive Element for Refractive Index Measuring Based on Dielectric Grating on a Metal Substrate

Volodymyr Fitio
Department of Electronic Engineering
Lviv Polytechnic National University
Lviv, Ukraine
volodymyr.m.fito@lpnu.ua

Iryna Yaremchuk
Department of Electronic Engineering
Lviv Polytechnic National University
Lviv, Ukraine
iryna.y.yaremchuk@lpnu.ua

Tetiana Bulavinets
Department of Electronic Engineering
Lviv Polytechnic National University
Lviv, Ukraine
tetiana.o.bulavinets@lpnu.ua

***Abstract* — The results of the numerical simulation of plane wave diffraction by the structure of the dielectric grating on a metal substrate are presented. In such a structure, the reflection coefficient is zero when the waveguide modes resonance (for TE and TM polarization waves) or the surface plasmon-polariton resonance for TM polarization occurs.**

***Keywords* — sensor, grating, resonance, waveguide mode, surface plasmon polariton**

I. Introduction

In work [1] it was shown for the first time that a zero-reflection coefficient can be achieved in the structure of a dielectric gating on a metal substrate due to field resonance, with carefully selected parameters of the grating. The reflection coefficient is zero under resonance of waveguide modes for waves of TE and TM polarizations or/and resonance of surface plasmon-polariton waves only for waves of TM polarization. The results of the resonance study in such structures as narrow-band absorbers of optical radiation were given in many works, particularly in [2 – 5]. Moreover, some of these works show that such structures can be used as sensors sensitive elements for measuring the refractive indices of liquids and gases [4, 5]

Dielectric gratings on dielectric substrates can be used as sensor's sensitive elements in which the resonance of waveguide modes occurs and the reflection coefficient from such structure is equal to unity [6, 7]. In addition, sensors for the refractive index measuring have been created based on a metal grating on a metal substrate [8 – 10]. The resonance of surface plasmon-polariton waves occurs at certain gratings parameters and the reflection coefficient from such structure is zero under resonance. Sensors have also been created based on a prism structure in which the surface plasmon-polariton resonance occurs and the reflection coefficient can be zero under resonance [11].

It should be noted that a large number of publications have recently been devoted to sensors for measuring the refractive index based on resonant phenomena of the electromagnetic field in the optical range, as evidenced by reviews on this issue [12,13], and the review [12] has a thousand references. In ref. [14], comparison of the characteristics of the sensor's sensitive elements based on a prism structure, a dielectric grating on a dielectric substrate, and a metal grating was made based on the waveguide phenomena theory. Analytical expressions relating the sensitivities of the sensors and their other characteristics to the change in the propagation constants of waveguide modes and surface plasma-polariton waves under the influence of the change in the testing medium refractive index have been obtained.

As already mentioned at the beginning of this section, field resonance is possible by the dielectric grating system on a metal substrate, which leads to total absorption [1]. It should be expected that when the refractive index of the medium surrounding the grating is changed, the resonance will be disrupted and it can be restored at a different wavelength or a different angle of incidence of the optical wave. Therefore, the study of such a structure from the point of view of its application as a sensitive sensor element is important and desirable.

II. Researched structure

The studied structure is shown in Fig. 1.

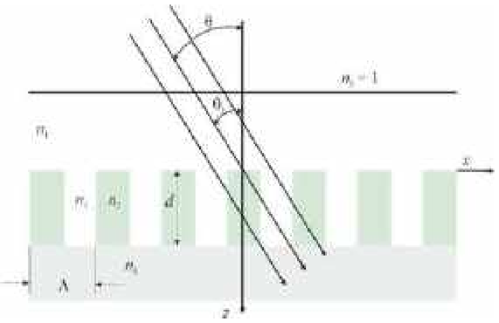


Fig. 1. Sheme of the dielectric grating on a metal substrate where θ is the angle of incidence of a plane wave in the air, θ_1 is the angle of incidence of the beam on the grating in the tested medium with the refractive index n_1, n_2 is the refractive index of the dielectric, n_3 is the refractive index of the metal (silver), d is grating thickness, Λ is the grating period.

Numerical modelling was carried out for the two indices of refraction (n_2) of the rectangular grating groove 1.47 and 2.0. Aqueous solutions were used as the research medium with the refractive index $n_1 = 1.333$ at the wavelength $\lambda = 1.064\ \mu\text{m}$. Therefore, we can see that the numerical analysis was carried out with a small and high contrast of the change of the grating dielectric constant. The grating filling factor is $F = 0.5$. In addition, calculations were carried out for two incidence angles of a plane wave on a periodic structure $\theta = 0$ and $\theta = \pi/18$. The diffraction calculation was carried out by rigorous coupled wave analysis (RCWA) [14, 15] using 41 coupled waves, which provided high calculation accuracy. The spectral dependences of the dielectric permittivity of silver in analytical form from work [16] were used for the numerical analysis.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

III. Results of the Numerical Analysis

The approximate values of the period Λ_0 and the grating thickness d_0 were determined using the theory of planar waveguides according to the rule given in [1] at the normal incidence of the plane wave on a periodic structure for the based parameters $\lambda_0 = 1.064$ and $n_{10} = 1.333$. At the next stages in the process of RCWA diffraction analysis, a reflection coefficient equal to zero was achieved by successive changes in the period of the grating and its thickness. The grating period and thickness are defined by Λ_{rez} and d_{rez}, respectively. The initial value of the grating thickness $d_0$was assumed to be equal to d_{rez} at normal incidence when a plane wave is incident at the angle $\theta = \pi/18$. The initial value of the grating period Λ_0 at the angle of incident θ was determined as follows:

$$\frac{2\pi}{\Lambda_0} = \frac{2\pi}{\Lambda_{rez}} \pm \frac{2\pi}{\lambda_0}\sin\theta, \quad (1)$$

where Λ_{rez} is the resonance grating period at the normal incidence of a plane wave. Based on Equation (1), when choosing the "+" sign, after simple transformations, Λ_0 can be obtained from the following expression:

$$\Lambda_0 = \frac{\lambda\Lambda_{rez}}{\lambda + \Lambda_{rez}\sin\theta}. \quad (2)$$

The "+" sign in Equation (2) provides only zero diffraction order of the reflected beam from the grating. The results of finding the approximate values of d_0, Λ_0, d_{rez}, and Λ_{rez} at the normal incidence of a plane wave and at the angle $\theta = \pi/18$ are summarized in Tables I and II for TE and TM polarizations, respectively.

TABLE I. GRATINGS PARAMETERS AT THE $\theta = 0$

No	n_2	Polar	d_0, nm	d_{rez}, nm	Λ_0,nm	Λ_{rez},nm	R_0
1	1.47	TE	400.9	921.8	777.5	789.49	$< 10^{-4}$
2	2.0	TE	938.7	753.1	696.6	709.4	$< 10^{-4}$
3	2.0	TM	675.8	717.5	750.2	730.13	$< 10^{-5}$
4	2.0	TM	61.25	75.75	750.2	732.56	$< 10^{-4}$

TABLE II. GRATINGS PARAMETERS AT THE $\theta = \pi/18$

No	n_2	Polar	d_0, nm	d_{rez}, nm	Λ_0,nm	Λ_{rez},nm	R_0
1	1.47	TE	921.8	923.7	699.38	698.28	$< 10^{-4}$
2	2.0	TE	753.1	755.7	635.8	607.7	$< 10^{-4}$
3	2.0	TM	717.5	739.0	632.4	658.00	$< 10^{-5}$
4	2.0	TM	75.75	100.0	654.3	644.13	$< 10^{-4}$

Analysis of Table 1 shows that the predicted grating thickness d_0 can be quite different from d_{rez} at $\theta = 0$. At the same time, Λ_0 and Λ_{rez} are quite close to each other. However, the predicted and resonance characteristics differ little from each other at $\theta = \pi/18$. Therefore, the approach of determining d_0 and Λ_0 according to Equation (2) provides sufficient accuracy and provides a quick search for d_{rez} and Λ_{rez}.

Thus, there is total absorption by the periodic structure. Changing any parameter of the periodic structure, the wavelength or the angle of incidence of the beam on the structure will lead to a disturbing of resonance. As a result, the reflection coefficient R_0 will becomes higher than zero, and may become close to unity. The most informative characteristics are the spectral dependencies, which are shown in Fig. 2. As can be seen from Fig. 2 the full width at half maximum (FWHM) is within wide limits for different cases. The FWHM for TE polarization waves is significantly lower than for TM polarization waves.

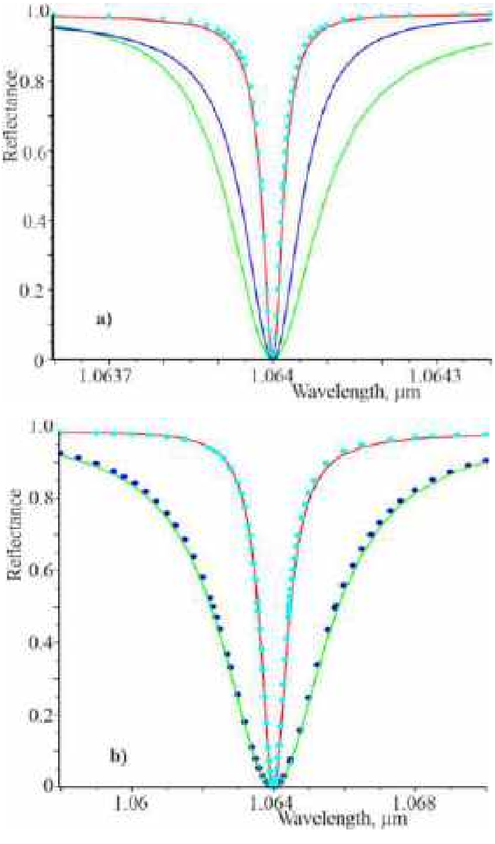


Fig. 2. Spectral dependences of the reflection coefficient. The red and green curves are obtained at θ = 0, and the blue and dark blue dots at θ $\theta = \pi/18$: a) is TE polarization, the red curve and blue dots correspond to the first lines of Table I and Table II, green curve and dark blue dots correspond to the second lines of Table I and Table II; b) is TM polarization, red curve and blue dots correspond to the third lines of Table I and Table II, green curve and dark blue dots correspond to the fourth lines of Table I and Table II

An important parameter of sensors is their sensitivity and two types of sensitivity are distinguished. The sensitivity S_λ is defined as the ratio of the change in the resonant wavelength to the change in the refractive index of the testing medium at the constant angle of incidence. The second one S_θ is defined as the ratio of the change in the resonant angle of incidence to the change in the refractive index of the testing medium at the constant wavelength. In our numerical studies, the sensitivity S_λ was determined as follows:

$$S_\lambda = \frac{\lambda_{rez}(n_1+\delta n_1) - \lambda_{rez}(n_1-\delta n_1)}{\delta n_1}, \quad (3)$$

where is $\delta n_1 = 0.0001$.

Knowing S_λ and FWHM, it is possible to calculate a figure of merit (FOM), which is an important characteristic of sensors, according to the expression:

$$\mathrm{FOM} = \frac{S_\lambda}{\mathrm{FWHM}}. \quad (4)$$

Calculated S_λ,FWHM and FOM for investigated structures are presented in Table III and Table IV.

TABLE III. PARAMETERS OF THE SENSORS ELEMENTS AT THE $\theta = 0$

No	n_2	Polar	d_{rez}, nm	$\Lambda_{rez,}nm$	s_λ, nm/RIU	FWHM, nm	FOM, RIU^{-1}
1	1.47	TE	921.8	789.49	553	0.045	12290
2	2.0	TE	753.1	709.4	470	0.200	2350
3	2.0	TM	717.5	730.13	400	0.925	432
4	2.0	TM	75.75	732.56	610	3.59	170

Where RIU is Refractive Index Unit

TABLE IV. PARAMETERS OF THE SENSORS ELEMENTS AT THE $\theta = \pi/18$

No	n_2	Polar	d_{rez}, nm	$\Lambda_{rez,}nm$	s_λ, nm/RIU	FWHM, nm	FOM, RIU^{-1}
1	1.47	TE	923.7	698.28	461	0.0404	11411
2	2.0	TE	755.7	607.7	269	0.125	2152
3	2.0	TM	739.0	658.00	400	0.892	448
4	2.0	TM	100	644.13	515	3.47	148

The maximum sensitivity S_λ is achieved for TM waves (fourth lines of the tables Table III and Table IV). At the same time, the lowest FWHMs are obtained for the first rows of these tables, which leads to the highest FOM values. The highest and the lowest FOM values differ by almost two orders of magnitude. In general, it can be stated that sensors based on resonance excitation for TE polarization waves have better characteristics compared to sensors based on resonance for TM polarization waves. At the same time, the lowest FWHM and, accordingly, the highest FOM are characteristic of TE waves of polarization at $n_2 = 1.47$, which coincides with the results of the work [1]. This is because when the contrast of the refractive index in the grating decreases, the FWHM decreases.

Therefore, it is interesting to obtain the dependence of the reflection coefficient from the grating R_0 on the change in the refractive index of the tested medium n_1. The corresponding dependences for the first line of Table IV (TE polarization) and the third line of the same table (TM polarization) are shown in Fig. 3. Within the limits of linear dependence, sensitivity can be calculated according to the following expression:

$$S = \frac{\delta R_0}{\delta n_1} = 13700 \text{ RIU}^{-1}. \qquad (5)$$

The change $\delta R_0 = 0.00137$ on the linear section when the refractive index changes $\delta n_1 = 10^{-7}$, it can be recorded by modern measuring equipment. In general, it can be stated that as a sensor-sensitive element is rational to use the resonance of waveguide modes for TE polarization waves with a small contrast of the dielectric constant of the grating.

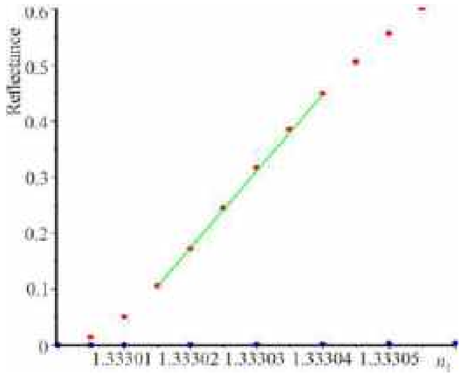


Fig. 3. Dependence of the reflection coefficient R_0 on the change in the refractive index of the tested medium n_1: red dots are TE polarization, green line are TM polarization, green line is linear approximation of the dependence of R_0 on n_1.

The field distribution at $z = 0$ (red color of curves) and at $z = d$ (blue color of curves) under waveguide modes resonance presented in Fig. 4, under resonance of surface plasmon-polariton waves presented in Figs. 5, 6.

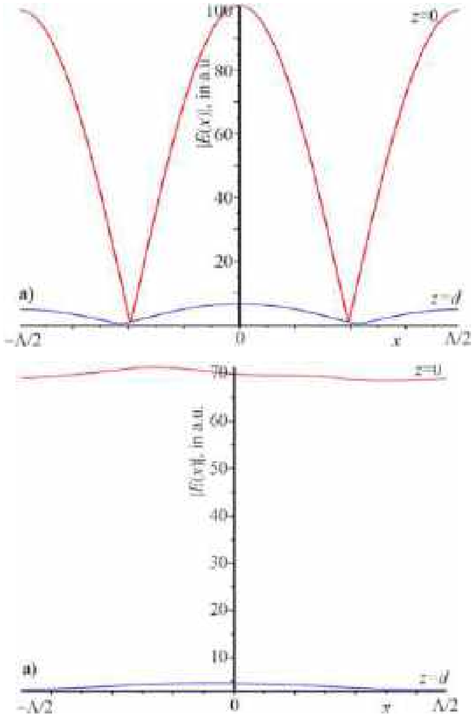


Fig. 4. Dependence of the electric field intensity modulus |E(x)| along the grating period for TE polarization waves (first lines of Table I and Table II): a) θ=0, b) θ=π/18.

It is known that the maximum field strength occurs at the metal/dielectric interface for a surface plasmon-polariton wave. There is the strongest field at $z = d$following Fig. 6, that is, at the metal/dielectric interface. Therefore, it can be asserted that only surface plasma-polariton waves can be excited at small grating thicknesses.

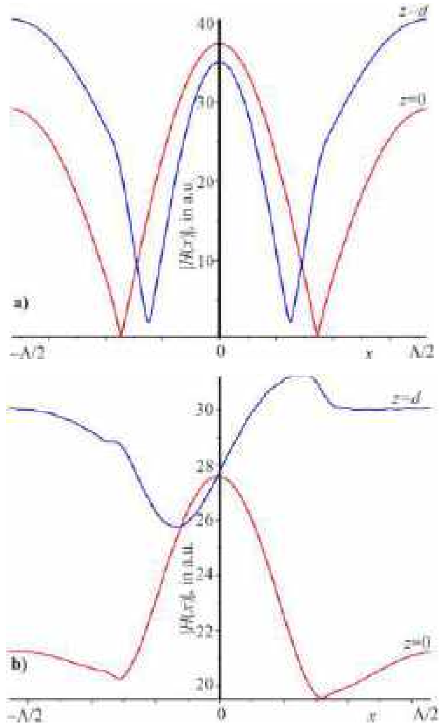


Fig. 5. Dependence of the magnetic field intensity modulus |H(x)| along the grating period for TM waves of polarization (third lines of Table I and Table II): a) θ=0, b) θ=π/18.

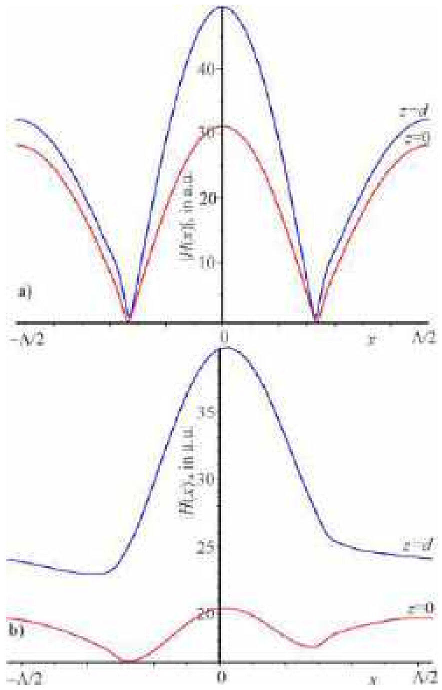


Fig. 6. Dependence of the magnetic field intensity modulus |H(x)| along the grating period for TM waves of polarization (fourth lines of Table I and Table II): a) θ=0, b) θ=π/18.

Both waveguide modes and surface plasmon-polariton waves can be excited with large grating thicknesses. In this case, the field can be higher at $z = 0$, and not at $z = d$, as shown in Fig. 4 for TE polarization waves, for which only waveguide modes can be excited by the grating.

Conclusion

Sensors sensitive element for refractive index measuring based on the dielectric grating on a metal substrate was theoretically studied. The maximum sensitivity of such a structure is achieved under the waveguide modes resonance with a small difference between the dielectric constant of the tested medium and the average value of the dielectric constant of the grating.

References

[1] V.M. Fitio, and Y.V. Bobitski, “Resonance effects in a dielectric grating; total absorption of electromagnetic waves by dielectric grating on the metal system”, J. Opt. A: Pure Appl. Opt., vol. 6, no 10, pp. 943 – 951, 2004.

[2] X. He, J. Jie, J. Yang, Y. Han, and S. Zhang, “Asymmetric dielectric grating on metallic film enabled dual-and narrow-band absorbers”, Optics Express, vol. 28, no 4, pp.4594-4602, 2020.

[3] X. He, Jinliang J., J. Yang, Y. Han, and S. Zhang, “Using fine-structured gratings to implement mid-infrared dual-band absorbers”, Eur. Phys. J. Appl. Phys., vol. 91, no 2, pp.20501, August 2020.

[4] M. Pan, H. Huang, W. Chen, S. Li, Q. Xie, F. Xu, and D. Wei, “Design of narrow-band absorber based on symmetric silicon grating and research on its sensing performance”, Coatings, vol.11, no 5, pp.553-561, May 2021.

[5] X. He, J. Jie, J. Yang, Y. Han, and S. Zhang, “Metal grating”, Applied Sciences,vol. 9, no 23, pp. 5022-5029, November 2019.

[6] N. Destouches, J.-C. Poimmer, O. Parriaux, T. Clausnitzer, N. Lyndin, and S.Tonchev, “Narrow-band resonant grating of 100% reflection under normal incidence”, Optics Express, vol. 14, no 26, pp. 12613-12622, December 2006.

[7] T. Tamulevičius, R. Šeperys., M. Andrulevičius, and S. Tamulevičius, “Total internal reflection based sub-wavelength grating sensor for the determination of the refractive index of liquids” Photonics and Nanostructures-Fundamentals and Applications, vol. 9, no 2, pp. 140-48, February 2011.

[8] K.H. Yoon, M.L. Shuler, and S.J. Kim, “Design optimization of nano-grating surface plasmon resonance sensors”, Optics Express, vol. 14, no 11, pp. 4842 – 4849, 2006.

[9] J. González-Colsa, G. Serrera, J m. Saiz, F González, F. Moreno, and P. Albella, “On the performance of a tunable grating-based high sensitivity unidirectional plasmonic sensor”, Optics Express, vol. 29, no 19, pp. 137332 – 13745, April 2021.

[10] S. Bellucci, O. Vernyhor, A. Bendziak, I. Yaremchuk, V. Fitio, Y. Bobitski, “Characteristics of the surface plasmon-polariton resonance in a metal grating, as a sensitive element of refraction index change”, Materials MDPI, vol. 13, no 8, pp. 1882-1891, April 2020.

[11] Y. Zeng, R. Hu, L. Wang, D. Gu, J. He, S.Y. Wu, et all. “Recent advances in surface plasmon resonance imaging: detection speed, sensitivity, and portability”, Nanophotonics, vol.6, no 5, pp.1017–1030, June 2017.

[12] Y. Xu, P. Bai, Xiaodong Z., Yu. Akimov, C. E. Png, L.-K. Ang, W. Knoll, and L. Wu, “Optical Refractive Index Sensors with Plasmonic and Photonic Structures: Promising and Inconvenient Truth”, Advanced Optical Materials, vol. 7, pp.1801433(47), 2019.

[13] G. Quaranta, G. Basset, O.J. F. Martin, and B.Gallinet, “Recent Advances in Resonant Waveguide Gratings”, Laser Photonics Rev., vol. 12, pp.1800017 (31) , 2018.

[14] M. G. Moharam, and T. K. Gaylord, “Rigorous coupled-wave analysis of metallic surface-relief grating”, J. Opt. Soc. Am. A.,vol. 3, no 5, pp. 1780-1787, 1986.

[15] L. Li, “Use of Fourier series in the analysis of discontinuous periodic structures”, J.Opt.Soc.Am. A.,vol. 13, no 9, pp. 1870–1876, 1996

[16] V. Fitio, I. Yaremchuk, O. Vernyhor, and Y. Bobitski, “Analytical expressions for spectral dependences of silver, gold, copper and aluminum dielectric permittivity”, Optica Applicata vol. L, no 2, pp.171-184, 2020.

Porous Layers as a Buffer for Synthesizing CdO/por-CdS/CdS Heterostructures

Yana Suchikova
The Department of Physics and Methods of Teaching Physics Berdyansk State Pedagogical University
Berdyansk, Ukraine
yanasuchikova@gmail.com

Sergii Kovachov
The Department of Physics and Methods of Teaching Physics Berdyansk State Pedagogical University
Berdyansk, Ukraine
essfero@gmail.com

Zhakyp Karipbaev
L.N. Gumilyov Eurasian National University
2 Satpayev Str.
Nur-Sultan, 010008, Kazakhstan
karipbayev_zht_1@enu.kz

Ihor Bohdanov
The Department of Physics and Methods of Teaching Physics Berdyansk State Pedagogical University
Berdyansk, Ukraine
naukabdpu@gmail.com

Anastasiia Lysak
Berdyansk State Pedagogical University, Berdyansk, Ukraine Institute of Physics, Polish Academy of Sciences,
Warsaw, Poland
https://orcid.org/0000-0002-3114-6526

Anatoli I. Popov
Institute of Solid State Physics University of Latvia
8 Kengaraga str.
LV-1063, Riga, Latvia
popov@latnet.lv

***Abstract* — This study presents the synthesis of the CdO/por-CdS/CdS heterostructure, where a porous CdS layer is employed as a buffer layer. The elemental composition of the structure was investigated using EDX surface mapping, where cadmium and sulfur exhibited uniform distribution, and cadmium oxide was predominantly detected on spheroidal crystallites. Raman spectroscopic analysis confirmed the presence of CdS and CdO crystalline phases, indicating the material's partial amorphousness. X-ray diffraction analysis revealed characteristic peaks for CdS and amorphous features in the structure, which might be associated with ultrafine crystallites and lattice deformations. The gathered data underscore the significance of the proposed method in synthesizing CdO/por-CdS/CdS heterostructures with porous layers serving as a buffer, paving the way for future research and the fabrication of materials for photocatalysis and magnetoelectronics.**

***Keywords* — buffer layer, electrochemical etching, electrochemical deposition, crystallites, heterostructures, cadmium oxide**

I. Introduction

In recent decades, nanomaterial science has emerged as one of the most rapidly evolving research fields due to nanostructures' myriad possibilities for technological and applied applications [1, 2]. This is particularly true for semiconducting and dielectric nanomaterials that can be harnessed in photovoltaics and other optoelectronic applications [3, 4].

Heterostructure materials have garnered intensive scrutiny in the contemporary scientific community owing to their distinctive properties and potential in diverse technological implementations [5, 6]. Various types of heterostructures have been considered, including but not limited to Ga_2O_3/GaAs [7], CdTe/CdSe [8], and CdSe/CdS [9]. Additionally, research has pivoted towards heterostructures based on oxide compounds [10, 11].

However, achieving the optimal quality of heterostructures often encounters challenges such as lattice mismatch or undesirable stresses at the interfaces between materials [12, 13]. One of the strategies to circumvent these challenges is employing an intermediary buffer layer [14, 15].

Utilizing porous layers as a "soft" underlayer is among the most promising strategies [16, 17]. Such underlayers can aid in stress alleviation and address lattice incompatibilities. Many methods exist for crafting such porous structures [18, 19]. While modern techniques like ion-track templates have been explored [20, 21], traditional methods like electrochemical etching remain relevant due to their simplicity and the feasibility of synthesizing large material batches [22-24].

Within the scope of this study, we propose a novel approach to synthesize the CdO/por-CdS/CdS heterostructure via electrochemical treatment of the CdS surface. We thoroughly analyze the heterostructure's morphological, compositional, and structural characteristics, facilitating a deeper understanding of its properties and potential applications.

II. Experiment

A. Samples for Experiment

Monocrystalline CdS samples with n-type conductivity of cubic symmetry were utilized and obtained via the Physical Vapor Transport (PVT) method. Before experimentation, the samples underwent a treatment regimen encompassing polishing using specialized pastes, followed by cleaning in alcoholic and acid solutions. The primary objective of this approach is to mitigate the influence of surface oxides on subsequent nanolayer formation.

B. Methodology and Apparatus

A procedure was developed for electrochemical deposition, encompassing multiple sequential stages to produce the heterostructures. The experiment was executed in a specialized three-electrode cell equipped with two working electrodes and an additional reference electrode, the silver chloride electrode EVL1M3. Current measurements were performed using a potentiostat, positioning the sample and platinum at a fixed distance of 1.5 cm. Additionally, the cell was equipped with auxiliary functional modules: an air blower and a mixer for uniform solution mixing. The CdO/por-CdS/CdS heterostructure was formed in three primary steps: anodic, cathodic electrochemical treatments, and chemical etching.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

The operating mechanism involves conducting anodic treatment in an electrolyte. Subsequently, the solution is saturated with acid, altering the position of the anode and cathode for cathodic processing. The final step necessitates switching off the cell's power and undergoing chemical treatment in the same electrolyte.

It's noteworthy to emphasize that such a detailed methodology ensures precise control over the heterostructure's properties, potentially leading to optimized performance in intended applications.

1) Anodic Processing

This stage aimed to form por-CdS. An electrolyte was used in the ratio of HNO_3:H_2O = 1:1 with a potential of 5V. The process duration was 3 minutes. During the electrochemical reaction at the anode, the material dissolves due to oxidative processes. This phenomenon is known as "anodic dissolution."

In the initial etching stage, there was a marked emission of bubbles both at the anode and cathode. These bubbles arose due to releasing oxygen and hydrogen from the electrolyte solution. This phenomenon might hinder the effective etching of the semiconductor surface. An active mixing technique of the solution was employed to minimize this effect. The used electrolyte was at room temperature, and the experiment was conducted under natural lighting. During the investigation, a pronounced increase in current density was observed, indicating the activity of electrochemical processes at the "electrolyte-semiconductor" interface. The color of the electrolyte solution turned yellow, which might indicate the presence of cadmium ions released from the crystal surface.

2) Cathodic Deposition

The primary goal of this stage was the formation of CdO. The process took place in the electrolyte solution of HNO_3:H_2O:C_2H_5OH in a ratio of 1.5:1:0.5. The potential for the reaction was set at 5V, with an action duration of 3 minutes.

The second stage of the experiment focused on the deposition of products from the electrochemical reaction onto the semiconductor. During the response at the cathode, a usual observation is the reduction process of ions, leading to the deposition of material on the cathode. This action is often called "cathodic deposition" or "electrolytic reduction". Unlike the first stage, there was no use of the mixing method for the electrolyte. Over 4 minutes, a gradual increase in current density was noted, after which its value became stable. Interestingly, the solution lightened, potentially indicating the active deposition of cadmium ions onto the crystal.

3) Chemical etching

During this stage, CdO formation also took place, but without using an electrical current. The reagent employed was a solution of HNO_3:H_2O:C_2H_5OH in a ratio of 1.5:1:0.5, and the procedure lasted for 1 minute. This action aimed to complete all electrochemical reactions and consolidate the surface characteristics of the semiconductor. After the experiment, the samples were thoroughly dried and stored in the open air for three months.

C. Characterization

Scanning electron microscopy was employed using the SEO-SEM Inspect S50-B device to study the morphological properties of the formed heterostructure. For analyzing the elemental composition of the sample surfaces, the EDX method was utilized, performing mapping and measurements at specific points. Structural features of the sample were examined using Raman spectroscopy with the RENISHAW inVia Reflex equipment. Parameters included a 5% laser intensity, an excitation wavelength of 532 nm, and a spectral range from 100 to 700 cm^{-1}.

III. Results

A. SEM Analysis

Figure 1 presents the SEM image that illustrates the morphology of the synthesized heterostructure of CdO/por-CdS/CdS. The structure can be observed to feature dispersed islands and spherical crystals. The surface layer demonstrates heterogeneity: some areas are predominantly fluffy, whereas others are covered in crystals. This surface configuration could suggest the presence of crystalline and amorphous regions in the topmost layer.

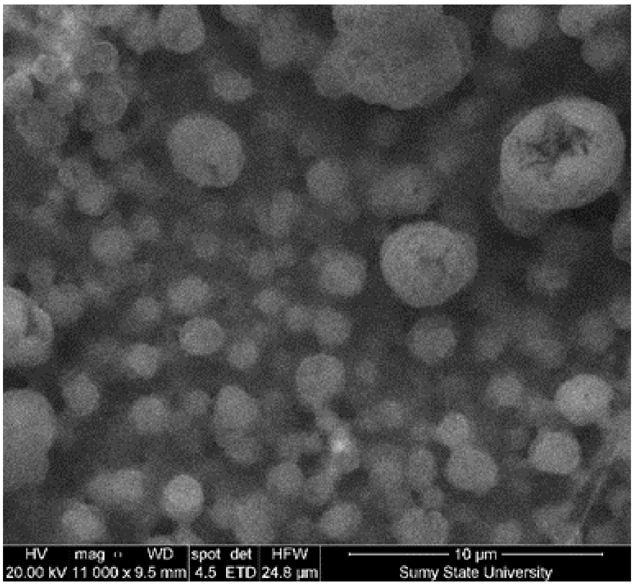


Fig. 1. SEM image showcasing the morphology of the synthesized CdO/por-CdS/CdS heterostructure.

The crystalline structures take on ring-like and spherical forms, with external diameters ranging from 0.5 to 5 µm. In those structures that adopt a ring-like shape, the internal diameter reaches up to 200 nm. Notably, rings with larger external diameters possess a more significant internal void, whereas the smaller ones may not have it. These findings can indicate a sequential formation of the structures: initially small spheres, which eventually grow in size, morphing into rings during subsequent synthesis stages.

Figure 2 depicts a cross-sectional view of the said structure. This layer stands out with its fluffy and porous texture. One can observe spheroidal crystallites filling up the spaces within these pores, imparting additional textural intricacies to the surface. These pores do not showcase a specific growth direction within the crystal. Instead, they form arbitrarily, creating massive etching pits. At the topmost portion, additional surface crystallites can be observed. The overall thickness of this layer approximates around 20 µm.

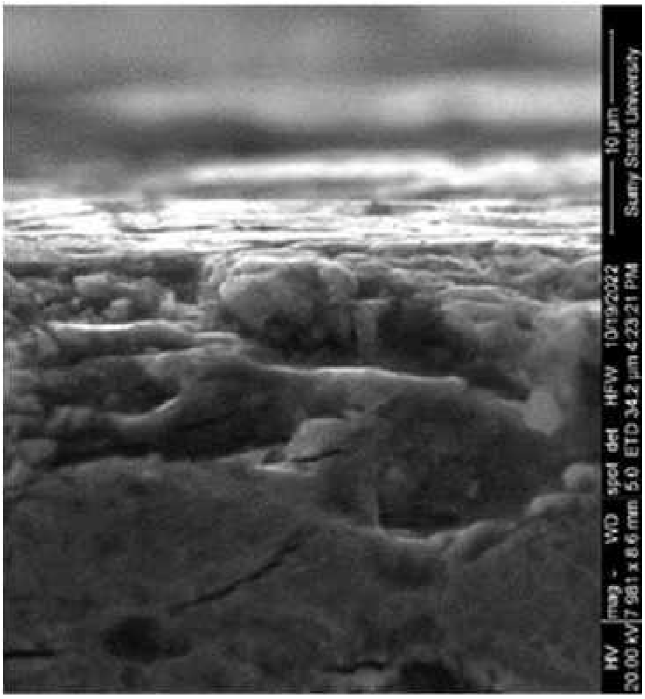


Fig. 2. Cross-sectional view of CdO/por-CdS/CdS heterostructure.

B. EDX Analysis

Figure 3 displays the outcomes of the EDX analysis carried out through the mapping technique for the sample's surface. The analysis reveals that cadmium and sulfur elements are uniformly distributed throughout the sample's surface. However, oxygen predominantly marks its presence on the spheroidal crystallites.

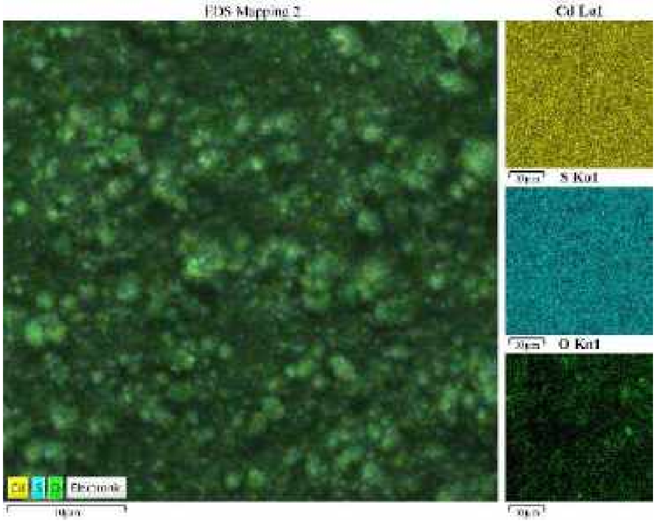


Fig. 3. Energy dispersive (EDX) mapping analysis of the surface CdO/por-CdS/CdS heterostructure.

The Electron Dispersive X-ray (EDX) spectroscopic analysis identified the surface's elemental composition at various points. The findings of the analysis are summarized in Table 1.

Several observations can be drawn from the data analysis in the table. The oxygen (O) content varies from 38.26% to 43.78%, with Point 3 exhibiting the maximum oxygen concentration. The sulfur (S) range sees a decline from Point 1 to Point 4, with the peak value at Point 2 (26.79%) and the lowest at Point 4 (15.27%). The cadmium (Cd) content seems relatively stable across the first three points, oscillating around 35%, but it escalates to 43.54% at Point 4. Such an elemental distribution implies some heterogeneity in the material or suggests that the formation processes or interactions between the components could have varied across different sections of the sample.

The average elemental composition of the surface consists of approximately 40.84% O, 22.015% S, and 37.145% Cd. Table 2 presents the ratio among the system's components.

TABLE I. ELEMENTAL COMPOSITION OF THE SURFACE AS DETERMINED THROUGH THE EDX METHOD

Element	At, %			
	1	2	3	4
O	40.13	38.26	43.78	41.19
S	24.68	26.79	21.32	15.27
Cd	35.19	34.95	34.9	43,54

TABLE II. RATIO OF COMPONENTS ON THE SURFACE CdO/POR-CdS/CdS HETEROSTRUCTURE

The ratio |	Value
Cd/S	1.686
Cd/O	0.909
S/O	0.538
Cd+S/O	1.447
Cd+O/S	3,543

The Cd/S ratio indicates that the number of Cd atoms is approximately 68.6% greater than that of S atoms. The Cd/O ratio is less than one, suggesting that the number of oxygen atoms slightly exceeds that of Cd atoms. The S/O ratio demonstrates that the number of oxygen atoms is twice as many as sulfur atoms. The ratios of Cd+S/O and Cd+O/S highlight significant structural heterogeneity, potentially pointing to various phases or regions of interaction among the components, corroborated by the EDX mapping data presented in Fig. 3. The observed variations in concentrations suggest that factors like reaction time, temperature, or reagent concentrations must be adjusted for a more homogeneous distribution during synthesis

C. Raman Analysis

Raman spectroscopy reveals peaks of high (462, 602 cm^{-1}), medium (253, 300 cm^{-1}), and low (112, 216 cm^{-1}) intensities (see Fig. 4). It is particularly noteworthy that certain spectral lines show expansion from the right side, with additional weak-intensity peaks also being present. The characteristic lines at 300, 462, and 602 cm^{-1} can be attributed to features of CdS, whereas peaks at 113, 216, and 253 cm^{-1} are correlated with the cubic structure of CdO. The critical peak at 602 cm^{-1} points to the 2LO mode of CdS, while 300 cm^{-1} is determined by the fundamental LO mode of CdS. This latter peak tends to have an asymmetric expansion, possibly due to nanoparticles with various sizes.

The line at 253 cm^{-1} characterizes cubic CdO. Notably, the lines in the range (100 – 250) cm^{-1} indicate the presence of cadmium oxide in the CdS structure. Their weak intensity compared to CdS peaks and noise and expansion might suggest partial crystallization of the surface layers, with the amorphous form of CdO likely prevailing.

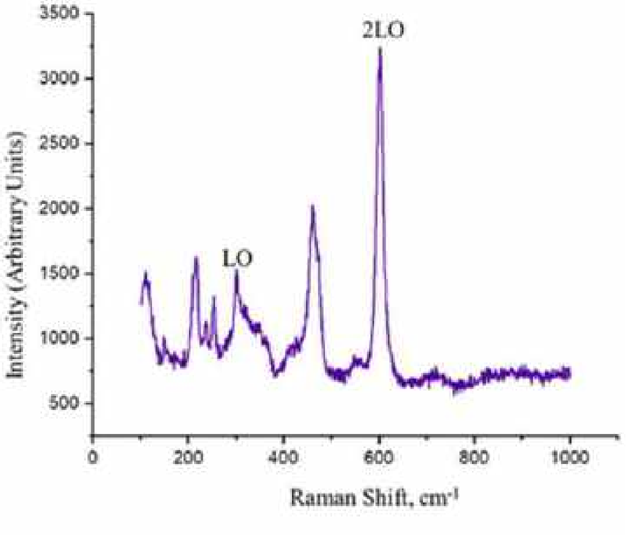


Fig. 4. Raman spectra of the CdO/por-CdS/CdS heterostructure.

D. XRD Analysis

Fig. 5 presents the diffractometric spectra of the obtained structure, compared with the standard ranges of CdS and CdO. A primary diffraction peak at 2θ = 26.4° is observed, which correlates with the (002) plane of the cubic phase of CdS, according to reference data PDF-4 00-001-0647. Other peaks in the plot have low intensity.

This spectral feature may have several reasons. Firstly, such “amorphization” of the structure could arise due to the presence of microscopic pores and ultrafine crystallites with sizes within the 2-5 nm range. This, in turn, may cause deformations in the crystalline lattice. Furthermore, the scattering of peaks and their low-intensity values might suggest significant crystallite size variability and irregularity variability. Additionally, the imperfect crystallization of the material may be a reason for the weak intensity of the spectrum, confirming the potential presence of both crystalline and amorphous phases in the cadmium oxide-based material.

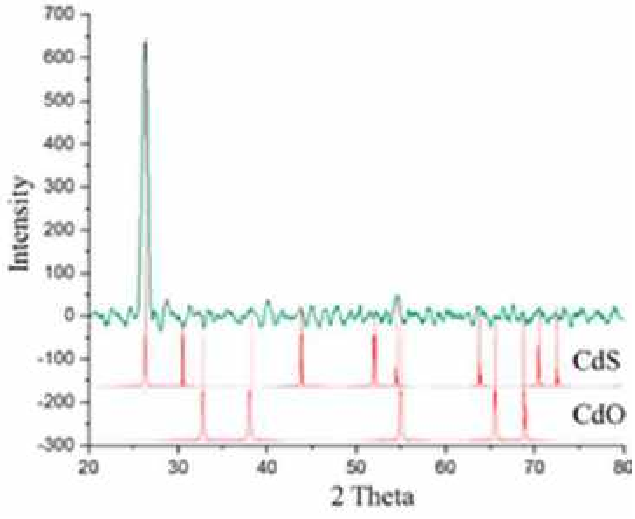


Fig. 5. XRD spectrum of CdO/por-CdS/CdS with overlaid reference spectra CdO, CdS, from the Crystallography Open Database (COD) visualized using the VESTA program.

IV. Discussion

Upon examining the research results on the newly formed structure based on CdS and CdO, several critical aspects of its formation and properties can be determined. The processes that occur during the electrochemical treatment of CdS play a role in shaping the unique structure of the material. This study demonstrates that the electrochemical etching of CdS leads to forming a porous layer, creating a surface density with numerous pits. This porous layer, combined with an electrolyte solution enriched with cadmium and sulfur atoms, facilitates the deposition of cadmium bound with oxygen during the subsequent anodic deposition stage. The porous structure ensures enhanced adhesion; as a result, crystalline growth is initiated at defect sites formed on the porous layer. Growth centers of crystallites include the peaks of porous islands where the cadmium oxide film undergoes recrystallization processes, developing spherical crystallites. Over time, these spheres transform into ring shapes with distinct crystalline characteristics, while the central part of the islands remains in an amorphous state. Due to electrochemical reactions, the material's surface acquires a new, porous structure, increasing its active surface for additional chemical processes. Such a structure dramatically enhances the efficiency of electrochemical reactions due to increased available surface area.

The similarities in the lattice characteristics and structural parameters of CdS and CdO underscore the importance of their coexistence in the formation of new nanostructures [25, 26]. The porous CdS may act as a "soft" substrate, reducing the risk of stress emergence in the overgrown CdO layer. Such an interfacial interaction between CdS and CdO can be a source of numerous applications, especially in high-tech sectors like photovoltaics.

The CdO layer, in turn, serves as a protective passivating layer for CdS, shielding it from unwanted external influences and ensuring the composite's stability. Oxide coatings, like CdO, typically have excellent passivating properties, protecting a more active material from corrosion and other aggressive environments [27, 28].

In the broader context, the newly formed structure showcases a range of intriguing characteristics that can be explored for future applications. Understanding its fundamental properties will aid in harnessing this structure for specific technological needs.

Conclusions

The research demonstrated the effectiveness of using porous layers as a buffer for synthesizing CdO/por-CdS/CdS heterostructures. The proposed method of electrochemical treatment of CdS, which involves sequential anodic etching to form por-CdS, cathodic deposition to form CdO on the surface of por-CdS, and final chemical etching, proved successful in obtaining the desired heterostructure.

Raman spectroscopic analysis of the heterostructure confirmed the presence of a volumetric CdS layer and amorphous and cubic phases of CdO. These data correlate with the EDX analysis results, which indicated a cadmium oxide film on the sample's surface.

Morphological analysis of the heterostructure using SEM revealed the presence of islands and ring-like crystallites. The recorded sizes of the crystallites and their shape attest to the success of the proposed self-assembly mechanism.

Based on the data obtained, it can be asserted that the proposed approach to the synthesis of CdO/por-CdS/CdS heterostructures, using porous layers as a buffer, may be significant for further research and development of new materials with potential applications in photocatalysis, magnetoelectronics, and other areas.

Acknowledgment

The study was supported by the Ministry of Education and Science of Ukraine via Project No. 0122U000129 "The search for optimal conditions for nanostructure synthesis on the surface of A3B5, A2B6 semiconductors and silicon for photonics and solar energy" and Project No. 0121U10942 "Theoretical and methodological bases of system fundamentalization of the future nanomaterials experts training for productive professional activity". In addition, the research was partly supported by COST Action CA20129 "Multiscale Irradiation and Chemistry Driven Processes and Related Technologies" (MultIChem), and COST Action CA20126 - Network for research, innovation, and product development on porous semiconductors and oxides (NETPORE). A.P. thanks to the Institute of Solid-State Physics, University of Latvia. ISSP UL as the Center of Excellence is supported through the Framework Program for European universities, Union Horizon 2020, H2020-WIDESPREAD-01–2016–2017-TeamingPhase2, under Grant Agreement No. 739508, CAMART2 project.

References

[1] A. Alfieri, S. B. Anantharaman, H. Zhang, and D. Jariwala, "Nanomaterials for quantum information science and engineering," Advanced Materials, p. 2109621, 2022.

[2] M. F. Hochella Jr, D. W. Mogk, J. Ranville, I. C. Allen, G. W. Luther, L. C. Marr, ..., and Y. Yang, "Natural, incidental, and engineered nanomaterials and their impacts on the Earth system," Science, vol. 363, no. 6434, p. eaau8299, 2019.

[3] H. Klym, I. Karbovnyk, A. Luchechko, Y. Kostiv, V. Pankratova, and A.I. Popov, "Evolution of free volumes in polycrystalline $BaGa_2O_4$ ceramics doped with Eu3+ ions," Crystals, vol. 11, no. 12, pp. 1515, 2021.

[4] A. Usseinov, Z. Koishybayeva, A. Platonenko, ..., Y. Suchikova, and A.I. Popov, "Ab-Initio Calculations of Oxygen Vacancy in Ga_2O_3 Crystals," Latvian Journal of Physics and Technical Sciences, vol. 58, no. 2, pp. 3–10, 2021.

[5] H. Klym, I. Karbovnyk, M.C. Guidi, O. Hotra, and A.I. Popov, "Optical and Vibrational Spectra of CsCl-Enriched GeS_2-Ga_2S_3 Glasses," Nanoscale Research Letters, vol. 11, no. 1, art. no. 132, 2016.

[6] V.P. Savchyn, A.I. Popov, O.I. Aksimentyeva, H. Klym, Y.Y. Horbenko, V. Serga, A. Moskina, and I. Karbovnyk, "Cathodoluminescence characterization of polystyrene-BaZrO3 hybrid composites," Low Temperature Physics, vol. 42, no. 7, pp. 760-763, 2016.

[7] Y. Suchikova, A. Lazarenko, S. Kovachov, A. Usseinov, Z. Karipbaev, and A. I. Popov, "Formation of porous Ga_2O_3/GaAs layers for electronic devices," in 2022 IEEE 16th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), February 2022, pp. 01-04.

[8] M. Molaei, F. Farahmandzadeh, T. S. Mousavi, and M. Karimipour, "Photochemical synthesis, investigation of optical properties and photocatalytic activity of CdTe/CdSe core/shell quantum dots," Materials Technology, vol. 37, no. 11, pp. 1818-1824, 2022.

[9] H. Li, C. Cheng, Z. Yang, and J. Wei, "Encapsulated CdSe/CdS nanorods in double-shelled porous nanocomposites for efficient photocatalytic CO_2 reduction," Nature Communications, vol. 13, no. 1, p. 6466, 2022.

[10] Y. Suchikova, S. Kovachov, I. Bohdanov, ..., V. Pankratov, and A.I. Popov, "Study of the structural and morphological characteristics of the $Cd_xTe_yO_z$ nanocomposite obtained on the surface of the CdS/ZnO heterostructure by the SILAR method," Applied Physics A: Materials Science and Processing, vol. 129, no. 7, pp. 499, 2023

[11] A. Singh, S. Sikarwar, A. Verma, and B. C. Yadav, "The recent development of metal oxide heterostructures based gas sensor, their future opportunities and challenges: a review," Sensors and Actuators A: Physical, vol. 332, p. 113127, 2021.

[12] A. Castellanos-Gomez, X. Duan, Z. Fei, H. R. Gutierrez, Y. Huang, X. Huang, ..., and P. Sutter, "Van der Waals heterostructures," Nature Reviews Methods Primers, vol. 2, no. 1, p. 58, 2022.

[13] E. Gabriel, C. Ma, K. Graff, A. Conrado, D. Hou, and H. Xiong, "Heterostructure engineering in electrode materials for sodium ion batteries: recent progress and perspectives," eScience, p. 100139, 2023.

[14] S. Kovachov, I. Bohdanov, Z. Karipbayev, ..., T. Tsebriienko, and A.I. Popov, "Layer-by-Layer Synthesis and Analysis of the Phase Composition of $Cd_xTe_yO_z$/CdS/por-ZnO/ZnO Heterostructure," in 2022 IEEE 3rd KhPI Week on Advanced Technology, KhPI Week 2022 - Conference Proceedings, 2022.

[15] S. Thalhammer, A. Hörner, M. Küß, S. Eberle, F. Pantle, A. Wixforth, and W. Nagel, "GaN Heterostructures as Innovative X-ray Imaging Sensors—Change of Paradigm," Micromachines, vol. 13, no. 2, p. 147, 2022.

[16] Y. Suchikova, S. Kovachov, I. Bohdanov, ..., A. Moskina, and A. Popov, "Characterization of $Cd_xTe_yO_z$/CdS/ZnO Heterostructures Synthesized by the SILAR Method," Coatings, vol. 13, no. 3, pp. 639, 2023.

[17] Y. Zhang, X. Zhu, and Y. Zhang, "Exploring heterostructured upconversion nanoparticles: from rational engineering to diverse applications," ACS nano, vol. 15, no. 3, pp. 3709-3735, 2021.

[18] Y. Suohikova, S. Vambol, V. Vambol, N. Mozaffari, and N. Mozaffari, "Justification of the most rational method for the nanostructures synthesis on the semiconductors surface," Journal of Achievements in Materials and Manufacturing Engineering, vol. 92, no. 1-2, 2019.

[19] E. Monaico, I. Tiginyanu, and V. Ursaki, "Porous semiconductor compounds," Semiconductor Science and Technology, vol. 35, no. 10, 103001, 2020.

[20] S. Ruiz-Gómez, C. Fernández-González, and L. Perez, "Electrodeposition as a tool for nanostructuring magnetic materials," Micromachines, vol. 13, no. 8, p. 1223, 2022.

[21] M. Kurniawan and S. Ivanov, "Electrochemically Structured Copper Current Collectors for Application in Energy Conversion and Storage: A Review," Energies, vol. 16, no. 13, p. 4933, 2023.

[22] S. Yana, "Porous indium phosphide: Preparation and properties," in Handbook of Nanoelectrochemistry: Electrochemical Synthesis Methods, Properties, and Characterization Techniques, 2016, pp. 283–306.

[23] S. A. Hasoon, I. M. Ibrahim, R. M. Al-Haddad, and S. S. Mahmood, "Fabrication of nanostructure CdS thin film on nanocrystalline porous silicon," International Journal of Current Engineering and Technology, vol. 4, no. 2, pp. 594-601, 2014.

[24] Y. A. Suchikova, V. V. Kidalov, and G. A. Sukach, "Influence of the carrier concentration of indium phosphide on the porous layer formation," Journal of Nano- and Electronic Physics, vol. 2, no. 4, pp. 75–81, 2010.

[25] G. Wang, L. Gong, Z. Li, B. Wang, W. Zhang, B. Yuan, and A. Kuang, "A two-dimensional CdO/CdS heterostructure used for visible light photocatalysis," Physical Chemistry Chemical Physics, vol. 22, no. 17, pp. 9587-9592, 2020.

[26] Kahane, S. V., Sasikala, R., Vishwanadh, B., Sudarsan, V., & Mahamuni, S. (2013). CdO–CdS nanocomposites with enhanced photocatalytic activity for hydrogen generation from water. International journal of hydrogen energy, 38(35), 15012-15018.

[27] E. J. Casey and C. L. Gardner, "Anodic Passivation by "CdO" Studied by ESR," Journal of The Electrochemical Society, vol. 122, no. 7, p. 851, 1975.

[28] E. Jang, S. Jun, Y. Chung, and L. Pu, "Surface treatment to enhance the quantum efficiency of semiconductor nanocrystals," The Journal of Physical Chemistry B, vol. 108, no. 15, pp. 4597-4600, 2004..

Optical Resonances and Enhancement of the Electric Fields in the Gap Between Two Spherical Metallic Nanoparticles

Andrii Korotun
Department of Information Security and Nanoelectronics
Zaporizhzhia Politechnic National University
Zaporizhzhia, Ukraine
Department of Metallic State Theory
G.V. Kurdyumov Institute for Metal Physics of the NAS of Ukraine
Kyiv, Ukraine
andko@zp.edu.ua

Garry Moroz
Department of Radio Engineering and Telecommunications
Zaporizhzhia Politechnic National University
Zaporizhzhia, Ukraine

Roman Korolkov
Department of Information Security and Nanoelectronics
Zaporizhzhia Politechnic National University
Zaporizhzhia, Ukraine

Igor Titov
Department of Information Security and Nanoelectronics
Zaporizhzhia Politechnic National University
Zaporizhzhia, Ukraine

***Abstract* — The paper studies the characteristics of the dipole optical nanoantennas which consist of two spherical metallic particles, separated by dielectric gap. The questions connected with the enhancement of the electric fields and with the excitation of the optical resonance are considered. The sizes of nanoparticles, at which the enhancement of the fields by the antenna is maximum, have been found. The frequencies, at which the maximum of the enhancement is reached, have been determined. An influence of material of nanoparticles and environment on the enhancement and frequencies of optical resonances has been analyzed.**

Keywords* — *spherical metallic nanoparticles, enhancement of the local electric fields, optical resonance, dielectric function, relaxation rate, dipole antennas

I. Introduction

Under the interaction of light with metallic nanoparticle, its conduction electrons can be controlled by the incident electric field of the wave in the collective oscillations, known as localized surface plasmonic resonance (SPR). It results in the sharp change in the incident radiation pattern and in such effects as subwavelength localization of electromagnetic energy, formation the high efficiency hot spots near the surface of the nanoparticles or the directed light scattering outside the structure. Localized SPR can also interact with the electromagnetic fields, emitted by molecules, atoms or quantum dots, located near nanoparticles, which, in turn, results in the strong modification of the emissive and non-emissive properties of the emitter. Since the localized SPR provides an effective transfer of electromagnetic energy from near to far field of metallic nanoparticles and vice versa, plasmonic nanostructures can be considered as nanoantennas since they operate similarly to radio antennas but at higher (optical) frequencies. Such nanoantennas support hybrid electromagnetic modes called plasmon-polaritons, which have extremely short wavelengths and demonstrate strong electromagnetic field confinement on the nanometer scale. In addition to their unique optical properties, the use of nanoantennas at optical frequencies is rather attractive for numerous applications such as ultrasensitive sensing [1,2], ultra-compact light-emitting devices [3–5], spectroscopy [6,7], visualization [8], energy harvesting and photodetection applications [9], particle collection [10], nonlinear optics [11], thermal ablation of malignant neoplasms [12] and many others. Therefore, the efforts of researchers have been focused on controlling and tuning the spectral properties of nanoantennas by modifying their geometry and sizes [13-18].

Antennas, that localize the field in the gap between two metallic nanoparticles, are called dimer antennas. They can usually be categorized into dipole antennas and "bow-tie" type antennas [19]. Dipole antennas are the most common radio antennas, so they have found their analogs in optics as well. It is known that the field near metallic structures can be enhanced due to plasmonic resonances and the scattering on an object with large surface curvature, in dipole antennas the enhancement occurs also in the gap between two nanoparticles. Such antennas can be used for the wide range of applications, such as directing of energy into subwavelength optoelectronic devices, and for fluorescence or Raman scattering enhancement [20, 21].

Let us point out that the question connected with the enhancement of the fields in the neighborhood of individual nanoparticles of different morphology was studied in the works [22-24]. At the same time, the study of the enhancement of the fields and excitation of optical resonances in dimeric structures has not been practically carried out. Hence, the problem connected with the study of the frequency dependencies for the enhancement of the fields in the gap between two spherical nanoparticles and with the influence of the sizes and material of the particles and the properties of surrounding dielectric on the enhancement is the actual problem.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

II. Basic Relations

Let us consider the question connected with the enhancement of the local electric fields in the gap between two spherical metallic nanoparticles with the radius R, located in the medium with the dielectric permittivity ϵ_{m} at the distance d from each other (Figure. 1). Such system is dipole nanoantenna.

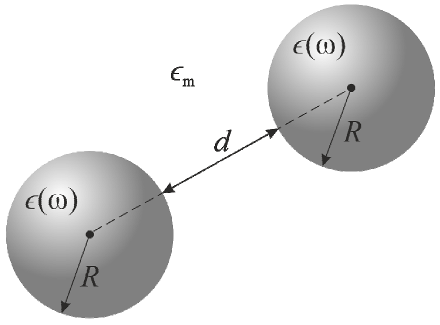


Fig. 1. Dipole nanoantenna of two spherical metallic nanoparticles in medium with the permittivity ϵ_{m}.

According to the definition, the enhancement of the field is equal to

$$\mathscr{G} = \frac{\mathscr{E}_{\max}}{\mathscr{E}_0}, \tag{1}$$

where $\mathscr{E}_{\max}$ is the maximum field, generated by the nanoparticles; $\mathscr{E}_0$ is the incident field amplitude.

Using the general expression for the fields under the presence of plasmonic nanoparticles and taking into account only the resonant term, we obtain [25]

$$\mathscr{E} \cong \frac{\epsilon_L \mathbf{e}_L \int\limits_V \left(\mathbf{e}_L \mathscr{E}_0\right) dV}{\left(\epsilon(\omega) - \epsilon_L\right) \int\limits_V \mathbf{e}_L^2 \, dV}, \tag{2}$$

where ϵ_L, $\mathbf{e}_L$ are the eigenvalue of the dielectric permittivity and eigenfunction of the electric field of resonant plasmon, and integration is performed over the volume of both nanoparticles; $\epsilon(\omega)$ is the dielectric function of material of the nanoparticles.

It is known that the longitudinal mode ($L = 1$) makes the dominant contribution into the local field. In this case the expression for the field enhancement has the form

$$\mathscr{G} = \frac{8}{3} \frac{R}{d} \left| \frac{1}{\epsilon(\omega) + 2\epsilon_{\mathrm{m}} \sqrt{R/d}} \right|. \tag{3}$$

Let us assume that Drude model is valid for the dielectric function. According to this model

$$\epsilon(\omega) = \epsilon^{\infty} - \frac{\omega_p^2}{\omega^2 + \gamma_{\mathrm{eff}}^2} + \mathrm{i}\frac{\omega_p^2 \gamma_{\mathrm{eff}}}{\omega\left(\omega^2 + \gamma_{\mathrm{eff}}^2\right)}, \tag{4}$$

where ϵ^{∞} is the contribution of ion core into the dielectric function of metal; $\omega_p = \sqrt{e^2 n_e / \epsilon_0 m^*}$ is the bulk plasmons frequency, e and n_e are the charge and concentration of electrons, correspondingly ($n_e = 3/4\pi r_s^3$, r_s is the average distance between electrons), ϵ_0 is the vacuum electric constant, and m^* is the effective mass of electrons.

An effective relaxation rate γ_{eff} is determined by the relation

$$\gamma_{\mathrm{eff}} = \gamma_{\mathrm{bulk}} + \gamma_{\mathrm{s}} + \gamma_{\mathrm{rad}}. \tag{5}$$

In formula (5) $\gamma_{\mathrm{bulk}} = \mathrm{const}$ is the bulk relaxation rate, and the surface relaxation rate and radiation attenuation rate are determined by the relations

$$\gamma_{\mathrm{s}} = \mathscr{A}(\omega, R) \frac{v_{\mathrm{F}}}{R}; \tag{6}$$

$$\gamma_{\mathrm{rad}} = \frac{2}{27} \mathscr{A}(\omega, R) \frac{v_{\mathrm{F}} R^2 \left(1 + 2\epsilon_{\mathrm{m}}\right)}{\sqrt{\epsilon_{\mathrm{m}}} \left(\epsilon^{\infty} + 2\epsilon_{\mathrm{m}}\right)} \left(\frac{\omega_p}{c}\right)^3, \tag{7}$$

where v_{F} is the Fermi velocity, and the effective parameter, which describes the degree of coherence loss under the electron scattering on the surface, has the form

$$\begin{aligned} \mathscr{A}(\omega, R) = \frac{3}{4} \frac{1}{1 + 2\epsilon_{\mathrm{m}}} \left(\frac{\omega_p}{\omega}\right)^2 \times \\ \times \left[1 - \frac{2\nu_s}{\omega} \sin\frac{\omega}{\nu_s} + 2\left(\frac{\nu_s}{\omega}\right)^2 \left(1 - \cos\frac{2\nu_s}{\omega}\right)\right], \end{aligned} \tag{8}$$

$\nu_s = v_{\mathrm{F}}/2R$ is the frequency of individual oscillations of electrons.

Taking into account

$$\epsilon(\omega) = \epsilon_1 + \mathrm{i}\epsilon_2,$$

where

$$\epsilon_1 = \epsilon^{\infty} - \frac{\omega_p^2}{\omega^2 + \gamma_{\mathrm{eff}}^2}; \tag{9}$$

$$\epsilon_2 = \frac{\omega_p^2 \gamma_{\mathrm{eff}}}{\omega\left(\omega^2 + \gamma_{\mathrm{eff}}^2\right)}, \tag{10}$$

we obtain the final expression for the fields enhancement in the gap between the nanoparticles from (3)

$$\mathscr{G} = \frac{8}{3} \frac{R}{d} \frac{1}{\sqrt{\left(\epsilon_1 + 2\epsilon_{\mathrm{m}} \sqrt{R/d}\right)^2 + \epsilon_2^2}}. \tag{11}$$

Let us point out that $\mathscr{G}(\omega)$ reaches its maximum

$$\mathscr{G}_{\max} = \frac{8}{3}\frac{R}{d}\frac{1}{\epsilon_2} \tag{12}$$

under the condition

$$\epsilon_1 + 2\epsilon_{\mathrm{m}}\sqrt{\frac{R}{d}} = 0\,, \tag{13}$$

hence, at the absence of the attenuation ($\gamma_{\mathrm{eff}} \to 0$), one can obtain the expression for the resonant frequency in non-dissipative approximation

$$\tilde{\omega}_{\mathrm{res}} = \sqrt{\frac{\omega_p^2}{\epsilon^\infty + 2\epsilon_{\mathrm{m}}\sqrt{R/d}}}\,. \tag{14}$$

In the case when the attenuation cannot be neglected, substituting (9) into (13), we obtain

$$\frac{\omega_p^2}{\omega_{\mathrm{res}}^2 + \gamma_{\mathrm{eff}}^2} = \epsilon_1 + 2\epsilon_{\mathrm{m}}\sqrt{\frac{R}{d}}\,. \tag{15}$$

Taking into account the relations (6) — (8), we obtain from (15)

$$\begin{gathered}\omega_{\mathrm{res}}^6 - \left[\frac{\omega_p^2}{\epsilon^\infty + 2\epsilon_{\mathrm{m}}\sqrt{R/d}} - \gamma_{\mathrm{bulk}}^2\right]\omega_{\mathrm{res}}^4 + \\ + 2\gamma_{\mathrm{bulk}}\mathscr{K}\omega_{\mathrm{res}}^2 + \mathscr{K}^2 = 0,\end{gathered} \tag{16}$$

where taking into account the smallness, compared to one, of the oscillating addends in square brackets of the formula (8)

$$\mathscr{K} = \frac{3}{4}\frac{\omega_p^2}{1+2\epsilon_{\mathrm{m}}}\frac{v_{\mathrm{F}}}{R}\left[1 + \frac{2}{27}\frac{1+2\epsilon_{\mathrm{m}}}{\sqrt{\epsilon_{\mathrm{m}}\left(\epsilon^\infty + 2\epsilon_{\mathrm{m}}\right)}}\left(\frac{\omega_p R}{c}\right)^3\right]. \tag{17}$$

As the third and the fourth addends in the equation (16) are small compared to the first two addends, we will solve it using the method of the successive approximations

$$\omega_{\mathrm{res}} = \omega_{\mathrm{res}}^{(0)} + \omega_{\mathrm{res}}^{(1)} + \ldots, \tag{18}$$

where

$$\omega_{\mathrm{res}}^{(0)} = \sqrt{\frac{\omega_p^2}{\epsilon^\infty + 2\sqrt{R/d}\,\epsilon_{\mathrm{m}}} - \gamma_{\mathrm{bulk}}^2}\,. \tag{19}$$

Substituting (18) into (16), we obtain

$$\omega_{\mathrm{res}}^{(1)} = -\frac{\mathscr{K}\left(\mathscr{K} + 2\gamma_{\mathrm{bulk}}\omega_{\mathrm{res}}^{(0)\,2}\right)}{2\omega_{\mathrm{res}}^{(0)}\left(\omega_{\mathrm{res}}^{(0)\,4} + 2\gamma_{\mathrm{bulk}}\mathscr{K}\right)}, \tag{20}$$

or finally for the frequency of the optical resonance

$$\omega_{\mathrm{res}} = \omega_{\mathrm{res}}^{(0)} - \frac{\mathscr{K}\left(\mathscr{K} + 2\gamma_{\mathrm{bulk}}\omega_{\mathrm{res}}^{(0)\,2}\right)}{2\omega_{\mathrm{res}}^{(0)}\left(\omega_{\mathrm{res}}^{(0)\,4} + 2\gamma_{\mathrm{bulk}}\mathscr{K}\right)}. \tag{21}$$

Hereafter we will use the relations (11) and (21) taking into account (5) — (10), (17) and (19) in order to obtain the numerical results.

III. THE RESULTS OF THE CALCULATIONS AND THEIR DISCUSSION

The calculations have been performed for dipole antennas, which consist of the nanoparticles of different metals with the different radiuses with the different distances between them in different mediums. The parameters of materials are given in Tables 1 and 2.

TABLE I. PARAMETERS OF METALS (SEE, FOR EXAMPLE, [26,27] AND THE REFERENCES IN IT)

Value	Metals					
	Al	Cu	Au	Ag	Pt	Pd
r_s / a_0	2.07	2.11	3.01	3.02	3.27	4.00
m^* / m_e	1.06	1.49	0.99	0.96	0.54	0.37
ϵ^∞	0.7	12.03	9.84	3.7	4.42	2.52
γ_{bulk}, 10^{14} s^{-1}	1.25	0.37	0.35	0.25	1.05	1.39

TABLE II. DIELECTRIC PERMITTIVITIES OF MATRICES (SEE, FOR EXAMPLE, [28] AND THE REFERENCES IN IT)

Value	Matrices					
	Air	CaF$_2$	Teflon	Al$_2$O$_3$	TiO$_2$	C$_{60}$
ϵ_{m}	1.0	1.54	2.3	3.13	4.0	6.0

Figure 2 shows the curves of the frequency dependencies for the enhancement in the gap between the particles of the fixed radius under the different distances between them (figure 2, *a*) and the particles of the different radiuses under the same distance (figure. 2, *b*). The results of the calculations show the presence of "red" shift $\mathscr{G}_{\max}$ under the decrease in gap between the particles and under the increase in radius of the particles with the constant distance between them. It should also be pointed out that "red" shift is accompanied by the increase in value of the field enhancement maximums, and the small-amplitude oscillations, caused by kinetic effects, are present in the infrared region ($\hbar\omega < 1$ eV). In turn, in the case of the gap $d = 10$ nm the enhancement for the particles with $R = 40$ nm is maximum.

Figure 3, *a* shows the results of the calculations $\mathscr{G}(\omega)$ for dipole antennas which consist of the particles of different metals. The indicated curves are similar to each other, but shifted in spectrum with respect to each other, because the values ω_{res} are significantly different for different metals. In this case, the field enhancement will be maximum in the gap between silver nanoparticles. It should also be pointed out that the properties of dielectric, in which the dimer antenna is located, also strongly influence the position and amplitude of the enhancement maxima. Thus, the decrease in permittivity ϵ_m results in “blue” shift of the enhancement maximums with the simultaneous increase in their value (Figure 3, *b*).

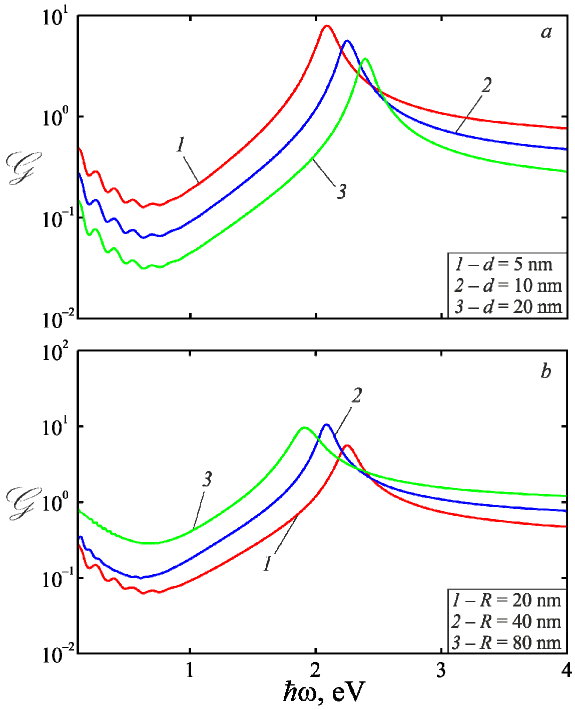


Fig. 2. The frequency dependence of the fields enhancement for the particles Au with the fixed radius under the different distances between them (*a*) and the particles with the different radius under the same distance (*b*)

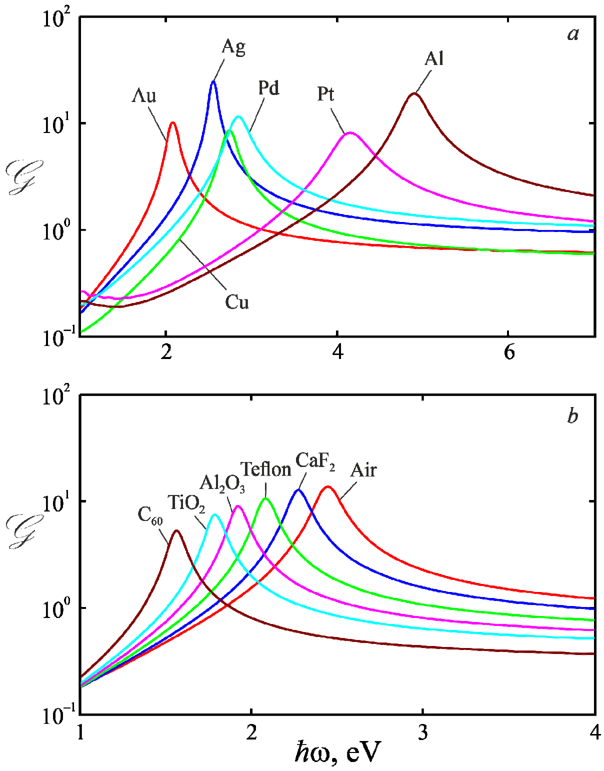


Fig. 3. The frequency dependence of the field enhancement for the particles of different metals in teflon (*a*) and particles Au in different dielectric mediums (*b*) at $R = 40$ nm , $d = 10$ nm .

Figure 4 shows the curves of the size dependence for the frequency of the optical resonance in the gap between two nanoparticles Au under the different distance between them. Let us point out that the curves $\omega_{res}(R)$ are qualitatively similar in the case of different d , and the resonant frequency decreases with the increase in size of the particles. The quantitative difference is that the resonant frequency for any radius of the particle will be greater for the greater distance between them.

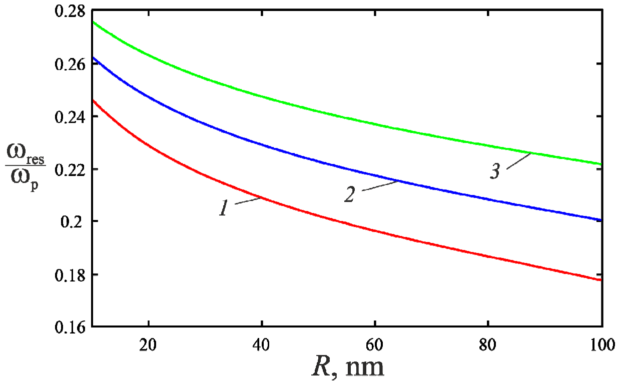


Fig. 4. The size dependence for the frequency of the optical resonanse in the gap between two nanoparticles Au with $R = 40$ nm under the different distance between them.

Conclusions

The relations for the frequency dependence of the fields enhancement in the gap between two spherical metallic nanoparticles and the relations for the size dependence of the frequency of the optical resonance have been obtained.

It has been demonstrated that the decrease in gap between the nanoparticles of the constant radius and the increase in radius of the nanoparticles under the constant distance between them result in the “red” shift of the fields enhancement maximum. At the same time, the value of the enhancement maximum increases with the decrease in distance between the particles, and the field enhancement for the nanoparticles with $R \sim 40$ nm is maximum in the case of the constant value of the gap.

It has been shown that if the system consists of the particles Au, Ag, Cu, then the enhancement maximums are reached in the optical frequency range, whereas in the case of the nanoparticles Pd, Pt Al are in the ultraviolet range, which is due to the significant differences between the optical parameters of these materials. Moreover, the enhancement is maximum in the case of dipole antenna which consists of two nanoparticles Au.

It has been established that the influence of the properties of the environment, surrounding the antenna, is reduced to the increase in amplitude of the enhancement maximums and in their “blue” shift under the decrease in value of dielectric permittivity.

The results of the calculations of the resonant frequencies indicate their decrease under the increase in radius of the nanoparticles which form the dimmer antenna.

The possibility of controlling the spectral position of the enhancement maximum by selecting the gap between the particles, the size and material of the nanoparticles and the materials of the dielectric medium, surrounding the antenna, has been demonstrated.

REFERENCES

[1] P. J. Schuck, D. P. Fromm, A. Sundaramurthy, G. S. Kino, and W. E. Moerner, “Improving the Mismatch between Light and Nanoscale Objects with Gold Bowtie Nanoantennas,” Phys. Rev. Lett., vol. 94, no 1, id. 017402, Jan. 2005.

[2] A. Konečná, M. K. Schmidt, R. Hillenbrand, and J. Aizpurua, “Probing the electromagnetic response of dielectric antennas by vortex electron beams,” Phys. Rev. Res., vol. 5, no. 2, id. 023192, Jun. 2023.

[3] V. Giannini, A. I. Fernández-Domínguez, S. C. Heck, and S. A. Maier, “Plasmonic nanoantennas: fundamentals and their use in controlling the radiative properties of nanoemitters,” Chem Rev, vol. 111, no. 6, pp. 3888–3912, Jun. 2011, doi: 10.1021/cr1002672.

[4] A. Nevet, N. Berkovitch, A. Hayat, P. Ginzburg, S. Ginzach, O. Sorias, and M. Orenstein, “Plasmonic nanoantennas for broad-band enhancement of two-photon emission from semiconductors,” Nano Lett, vol. 10, no. 5, pp. 1848–1852, May 2010.

[5] J. P. D. Brown Robert, Ed., Handbook of Optoelectronics: Concepts, Devices, and Techniques (Volume One), 2nd ed. Boca Raton: CRC Press, 2017.

[6] J. Aizpurua, G. W. Bryant, L. J. Richter, F. J. García de Abajo, B. K. Kelley, and T. Mallouk, “Optical properties of coupled metallic nanorods for field-enhanced spectroscopy,” Phys. Rev. B, vol. 71, no. 23, p. 235420, Jun. 2005.

[7] T. H. Taminiau, F. D. Stefani, F. B. Segerink, and N. F. van Hulst, “Optical antennas direct single-molecule emission,” Nat. Photonics, vol. 2, no. 4, pp. 234–237, 2008.

[8] S. Kawata, Y. Inouye, and P. Verma, “Plasmonics for near-field nano-imaging and superlensing,” Nature Photon, vol. 3, no. 7, Art. no. 7, Jul. 2009.

[9] M. W. Knight, H. Sobhani, P. Nordlander, and N. J. Halas, “Photodetection with Active Optical Antennas,” Science, vol. 332, no. 6030, pp. 702–704, May 2011.

[10] M. L. Juan, M. Righini, and R. Quidant, “Plasmon nano-optical tweezers,” Nature Photon, vol. 5, no. 6, Art. no. 6, Jun. 2011.

[11] T. Utikal, T. Zentgraf, T. Paul, C. Rockstuhl, F. Lederer, M. Lippitz, and H. Giessen, “Towards the Origin of the Nonlinear Response in Hybrid Plasmonic Systems,” Phys. Rev. Lett., vol. 106, no. 13, p. 133901, Mar. 2011.

[12] M. J. Rabienejhad, A. Mazaheri, and M. Davoudi-Darareh, “Design and optimization of nano-antenna for thermal ablation of liver cancer cells,” Chinese Phys. B, vol. 30, no. 4, p. 048401, Apr. 2021.

[13] E. Prodan, C. Radloff, N. J. Halas, and P. Nordlander, “A Hybridization Model for the Plasmon Response of Complex Nanostructures,” Science, vol. 302, no. 5644, pp. 419–422, Oct. 2003.

[14] P. Ginzburg, N. Berkovitch, A. Nevet, I. Shor, and M. Orenstein, “Resonances On-Demand for Plasmonic Nano-Particles,” Nano Lett., vol. 11, no. 6, pp. 2329–2333, Jun. 2011.

[15] J.B. Lassiter, H. Sobhani, J.A. Fan, J. Kundu, F. Capasso, P. Nordlander, and N.J. Halas, “Fano Resonances in Plasmonic Nanoclusters: Geometrical and Chemical Tunability,” Nano Lett., vol. 10, no. 8, pp. 3184–3189, Aug. 2010.

[16] N. Berkovitch and M. Orenstein, “Thin Wire Shortening of Plasmonic Nanoparticle Dimers: The Reason for Red Shifts,” Nano Lett., vol. 11, no. 5, pp. 2079–2082, May 2011.

[17] E. Üstün, Ö. Eroglu, U. M. Gür, and Ö. Ergül, “Investigation of nanoantenna geometries for maximum field enhancements at optical frequencies,” in 2017 Progress In Electromagnetics Research Symposium - Spring (PIERS), May 2017, pp. 3673–3680.

[18] G. Işıklar, Ş. Yazar, H. İbili, G. Onay, Z. Ahdab, and Ö. Ergül, “Computational design of nanoantennas with improved power enhancement capabilities via shape optimization,” Optical Engineering, vol. 62, Jan. 2023.

[19] A. E. Krasnok, I. S. Maksymov, A. I. Denisyuk, P. A. Belov, A. E. Miroshnichenko, C. R. Simovski, and Y. S. Kivshar, “Optical nanoantennas,” Physics-Uspekhi, vol. 56, no. 6, pp. 539–564, 2013.

[20] A. Sundaramurthy, K. B. Crozier, G. S. Kino, D. P. Fromm, P. J. Schuck, and W. E. Moerner, “Field enhancement and gap-dependent resonance in a system of two opposing tip-to-tip Au nanotriangles,” Phys. Rev. B, vol. 72, no. 16, p. 165409, Oct. 2005, doi: 10.1103/PhysRevB.72.165409.

[21] Alemayehu Nana Koya, M. Romanelli, J. Kuttruff, N. Henriksson, A. Stefancu, G. Grinblat, A. De Andres, F. Schnur, M. Vanzan, M. Marsili, M. Rahaman, A. Viejo Rodríguez, T. Tapani, H. Lin, B. Dalga Dana, J. Lin, G. Barbillon, R. Proietti Zaccaria, D. Brida, D. Jariwala, L. Veisz, E. Cortés, S. Corni, D. Garoli, and N. Maccaferri, “Advances in ultrafast plasmonics,” Appl. Phys. Rev., vol. 10, no. 2, p. 021318, Jun. 2023.

[22] S.A. Maier, “Plasmonics: Fundamentals and Applications,” New York: Springer, 2007, 224 p.

[23] K. Tanabe, “Field enhancement around metal nanoparticles and nanoshells: A systematic investigation,” J. Phys. Chem. C., vol. 112, no. 40, pp. 15721–15728, October 2008.

[24] A.V. Korotun, A.O. Koval, and V.V. Pogosov, “Optical parameters of bimetallic nanospheres,” Ukr. J. Phys., vol. 66, no. 6, pp. 518–527, July 2021.

[25] V. Klimov, Nanoplasmonics. New York: Jenny Stanford Publishing, 2014. doi: 10.1201/b15442.

[26] A. Korotun, N.Smirnova, V. Reva, I. Titov, “The spectral quality factor of the sensory elements of the nanosensors based on the surface plasmonic resonance,” 2021 IEEE 12th Int. Conf. on Electronics and Information Technologies, ELIT 2021 - Proceedings, 2021, pp. 216–221, May 2021.

[27] A.V. Korotun, Ya. V. Karandas, “Surface Plasmons in a Nanotube with a Finite-Thickness Wall,” Phys. Met Metallog., vol. 123, no. 1, pp. 7–15, January 2022.

[28] N.A. Smirnova, M.S. Maniuk, A.V. Korotun, and I.M. Titov “Optical absorption of the composite with the nanoparticles, which are covered with the surfactant layer,” Phys. Chem. Solid State, vol. 24, no. 1, pp. 181–189 2023.

Influence of Preparation Conditions on Structure and Properties of Composites Based on Polyethylene Oxide and Clay Nanoparticles

Eduard Lysenkov
Faculty of Computer Science
Petro Mohyla Black Sea National University
Mykolaiv, Ukraine
ealysenkov@ukr.net

Sergiy Bilyi
Department of Polymer Physics
Institute of Macromolecular Chemistry
NAS of Ukraine
Kyiv, Ukraine
sergeybilyi@gmail.com

Stanislav Nesin
Department of Polymer Physics
Institute of Macromolecular Chemistry
NAS of Ukraine
Kyiv, Ukraine
nesin@nas.gov.ua

Valeriy Klepko
Department of Polymer Physics
Institute of Macromolecular Chemistry
NAS of Ukraine
Kyiv, Ukraine
klepko_vv@ukr.net

Abstract — **The structure, thermophysical and electrical properties of polymer composite systems based on polyethylene glycol and clay nanoparticles were studied using the methods of X-ray structural analysis, differential scanning calorimetry and impedance spectroscopy. From the analysis of X-ray scattering data, it was found that the optimal degree of delamination of montmorillonite, which corresponds to the maximum interlayer distance, occurs in 3-5 minutes. A further increase in the mixing time has no significant effect on the structural characteristics of the composite. On the basis of calorimetric and impedancemetric studies, it is shown that the melting and glass transition temperatures, as well as the crystallinity and electrical conductivity, reach critical values at 3 min of treatment, after which they remain unchanged. It was established that the time of extruder mixing is optimal. At the same time, the maximum intercalation takes place, which leads to an impact on the final functional characteristics of the polymer-nanoclay system.**

Keywords* — *nanoclay, polymer composites, interlayer distance, thermophysical properties, crystallinity, electrical conductivity.

I. Introduction

Polymer nanocomposites are one of the most promising materials at the current stage of science and technology development. The growth of the fields of application of polymer nanocomposites is due to their unique physical and chemical properties [1]. Due to the combination of organic and inorganic components, these materials are characterized by increased strength, wear resistance, elasticity, etc. The combination of various fillers with a wide range of polymer matrices makes it possible to obtain materials with the necessary properties for high-tech areas such as energy, nanoelectronics, medicine, etc. [1-3]. Among the wide variety of fillers used to create polymer nanocomposites, nanoclays (montmorillonite, laponite, bentonite) deserve special attention due to their low cost, high strength and stability, efficiency when even a small amount is introduced [4]. In addition to other fields of application, materials based on polymer and nanoclay are promising for the creation of various electronic devices, sensors [5], polymer electrolytes [6] and coatings with high dielectric constant [7].

In modern scientific literature, much attention is paid to the composition and concentrations of components of polymer nanocomposites containing nanoclays. Also, it was established that the preparation conditions significantly affect the final properties of polymer-organoclay systems [8, 9].

The final properties of polymer nanocomposites depend significantly on the degree of stratification of the layered filler. However, obtaining a material with a high degree of intercalation or exfoliation is a very difficult task, because the energy that will hold the organoclay plates in the tactoid (pack) is very high [10]. To exfoliate or intercalate nanoclay tactoids, methods of ultrasonic dispersion or extruding are usually used. Mixing systems of the polymer-organoclay type using these methods leads to partial delamination of nanoclay plates [11]. For the intercalation of a polymer macromolecule in the interlayer space of a layered organoclay, it is necessary to apply a certain amount of energy, that is, to conduct mixing for a certain time. This energy depends on the type of polymer, its physical and chemical properties, as well as the type and properties of the nanoclay itself. For example, the effect of ultrasonic dispersion time on the structure and properties of polymer nanocomposites containing organoclay was investigated [12, 13]. It was established that the highest structural and thermophysical characteristics were observed after 10 minutes of ultrasonic treatment. Delva et al. [14] studied the effect of mixing energy on the properties of polymer nanocomposites based on polypropylene and montmorillonite by repeated extrusion. It is shown that the largest value of the interlayer distance of organoclay was observed at 12 cycles of extrusion, while the maximum of the modulus of elasticity was found at 7 cycles. The nature and content of nanoclay also influence on structure and functional properties of polymer nanocomposites based on polyethylene oxide [15]. It is established that in systems filled with montmorillonite, partial intercalation processes take place, while in systems filled by laponite complete exfoliation is observed. From the data of X-ray diffraction analysis it can be seen that the introduction of organoclay leads to amorphization of the system and promotes the formation of large crystallites.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

However, the number of such studies is limited, and this problem of establishing optimal conditions for the preparation of polymer-organoclay nanocomposites is urgent and requires more in-depth research. Therefore, the purpose of this work was to study the influence of preparation time on the structural, thermophysical and electrical characteristics of polymer nanocomposites using the polyethylene oxide-nanoclay system as an example.

II. Experimental part

A. Materials

Polyethylene oxide (PEO 1000), $HO[-CH_2-CH_2-O-]nH$ ($n \approx 22$) molecular weight $M_w = 1000$, produced by the Aldrich company, was chosen as the polymer matrix. At $T = 20$ °C, PEO-1000 is a solid substance with a density of $\rho = 1070$ kg/m^3. Melting point $T_m \approx 34$-35 °C.

As nanoclay (NC) filler montmorillonite from the Pizhev deposit was previously purified. Organomodified montmorillonite was obtained by treating the *Na* form of minerals with hexadecyltrimethylammonium bromide (manufactured by Merck) at a temperature of 75 °C for 24 hours.

B. Preparation

Before use, PEO was dehydrated by heating in a vacuum for 4 hours at 80-100 °C at a residual pressure of 300 Pa. The filler content in the polymer composite was 5 wt. %. (further %).

Polymer composite materials were produced by the method of extrusion (mechanical grinding in a melt) using a piston extruder followed by cooling according to room temperature. The main advantage of piston extruders over screw extruders is the ability to vary the time of mixing filler particles with a molten polymer matrix, after which the test sample can be formed in the form of either a plate or a thread. The manufacturing technology of the studied materials is given below [16]. The mixing time varied from 2 to 10 min.

C. Methods

The structure of composite at a small spatial scale was investigated using wide angle X-ray scattering (WAXS) instrument XRD-7000 (Shimadzu, Japan) with Cu K$_\alpha$ source of emission at a wavelength $\lambda = 0.154$ nm.

Thermophysical studies were performed in a dry atmosphere in the temperature range from -100 °C to 100 °C at a heating rate of 10 °C/min by DSC on the device DSC-60 Plus (Shimadzu, Japan). The midpoint method was used to determine the glass transition temperature, and the extremum point of the corresponding peak was chosen as the melting temperature.

Electrical conductivity of composite materials was studied using immitancemeter E7-20 in the frequency range $2.5\cdot10^1$-$1\cdot10^6$ Hz. A direct current conductivity was determined as value, when conductivity is frequency independent. Stainless steel electrodes were used for the research, the constant gap between the electrodes was 100 μm.

III. Results and discussion

To study the influence of preparation time on the formation and final properties of polymer composites based on polyethylene glycol and NC, structural features, thermophysical and electrical characteristics were studied.

A. Structural Features of the PEO-NC System

The effect of the filler on the structure of the polymer matrix in the range of sizes up to 4 nm was studied using WAXS method.

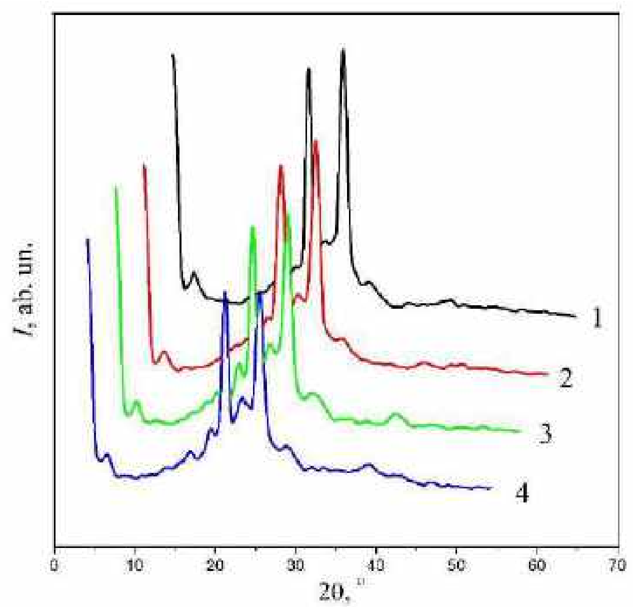


Fig. 1. Diffractograms of wide-angle X-ray scattering for polymer composites based on PEO and NC, produced by extrusion during: 1 – 2 min; 2 – 3 min; 3 – 5 min; 4 – 10 min.

Fig. 1 presents the WAXS data for PEO-NC composites (filler content was 5%) produced with different extrusion times. It is observed that the extrusion time significantly affects the structural properties of the composite. The graphs (Fig. 1) show a series of maxima. Peaks in the 5° region correspond to the presence of an ordered structure of NC [17]. From the parameters of this peak, it is possible to determine the interplane distance of NC. The interlayer distance of montmorillonite is the distance between two adjacent plates in the tactoid depends on the method of mixing the composite and the presence of organic modifiers [18]. The interplayer distance of montmorillonite (d) was determined using the Bragg equation [19]:

$$\lambda = 2d\sin\theta, \qquad (1)$$

where λ is the wavelength of characteristic X-ray radiation (λ = 1.54 Å), which was used in the study; θ is half the diffraction angle.

Fig. 3 presents dependence of the interlayer distance of NC on the extrusion time. The figure shows that the value of the interlayer distance of montmorillonite varies from 1.74 nm to 1.82 nm. The value of the interlayer distance is close to unintroduced montmorillonite 1.6-1.7 nm [18]. This fact indicates the insufficiency of the energy provided to the system to exfoliate the NC tactoids, so the polymer chains cannot fully intercalate in the interlayer space of the nanoclay. When the mixing time increases from 3 to 5 min, there is a sharp jump in the parameter d, which increases to a value of 1.82 nm. With this duration of processing, the system receives enough energy to partially destroy the montmorillonite tactoids, which leads to intercalation of the polymer in the interlayer space of the nanoclay. A further increase in the extrusion time does not lead to an increase in the interlayer distance. This indicates a certain energy saturation of the system, while the increase in energy absorbed by the polymer composite does not lead to the destruction of tactoids, but probably leads to the destruction of the polymer matrix.

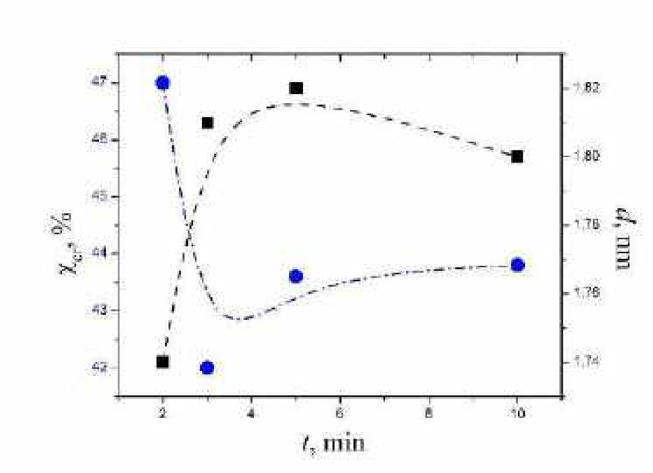


Fig. 2. Dependence of the structural characteristics of the polymer matrix (degree of crystallinity) and filler (interlayer distance) on the extrusion time for the investigated systems.

The most intense maxima observed in fig. 1 in the region of 20-30 degrees indicate the crystalline structure of the polymer component of the studied systems. If a crystalline phase is present in a polymer-containing system, its relative degree of crystallinity can be estimated. The determination of the relative degree of crystallinity (χ_{cr}) was carried out according to the method of Matthews [20], which is based on the comparison of the areas of the diffraction maxima characterizing the crystalline structure of the amorphous-crystalline polymer, (Q_{cr}), with the total area of the diffraction curve in the selected information angular interval ($2\theta_1$–$2\theta_2$):

$$\chi_{cr} = \frac{Q_{cr}}{(Q_{cr} + Q_{am})} \cdot 100\,\% \,. \tag{2}$$

Fig. 2 also presents the dependence of the relative degree of crystallinity of the polymer matrix on the extrusion time for PEO-NC systems. It is observed that the degree of crystallinity sharply decreases in the interval of processing times from 3 to 5 min, after which it almost does not change. This behavior correlates with the behavior of the structural characteristics of NC and is explained by the processes of intercalation of PEO macromolecules in the interlayer distance of montmorillonite. When the dispersion time increases to 5 min, the intercalation processes increase the surface of NC, which is able to contact with PEO. Due to these steric hindrances, polymer macromolecules cannot form crystalline structures upon cooling from the melt during the formation of the composite. Thus, the degree of crystallinity of the system decreases.

B. Thermophysical Characteristics of the PEO-NC System.

The processes of intercalation of PEO macromolecules in the interlayer simple NC significantly affect the functional characteristics of polymer composites, in particular, the thermophysical properties. Fig. 3 presents differential scanning calorimetry data for PEO-based polymer nanocomposites in the temperature range from –100 to 100 °C.

For all composites, two temperature transitions are observed on the DSC curves: glass and melting transitions. The glass process takes place in the temperature range of –81 to –76 °C. Intense maxima in the temperature range from 15 to 50 °C indicate the melting of the crystalline phase of PEO.

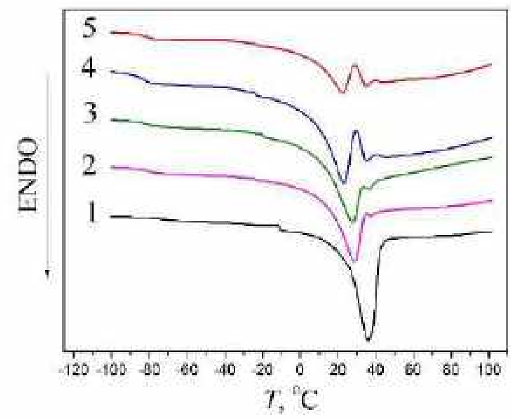


Fig. 3. DSC curves for polymer nanocomposites based on PEO and NC, produced by extrusion during: 1 – 0 min; 2 – 2 min; 3 – 3 min; 4 – 5 min; 5 – 10 min.

Different conditions for the preparation of systems of the polymer-organoclay type significantly affect the characteristics of temperature transitions of nanocomposite systems [18]. Fig. 4 shows the dependence of the glass transition temperature (T_g) on the extrusion time for the investigated systems.

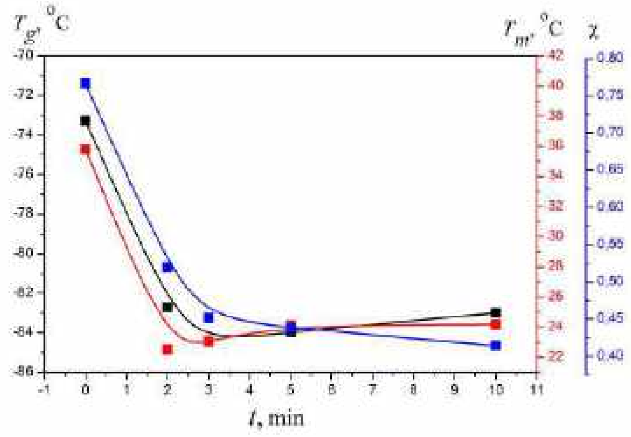


Fig. 4. Dependence of the glass transition and melting temperatures as well as degree of crystallinity on mixing time for polymer nanocomposites based on PEO and NC.

The glass transition temperature decreases with an increase in the processing time with the help of an extruder (Fig. 4). It reaches a minimum at the time of 3 minutes, after which it almost does not change. Therefore, the processes of intercalation and growth of the total surface area of the filler in the polymer composite significantly affect the cooperative movement of PEO macromolecules. This movement becomes more difficult with increasing processing time due to the steric hindrances that the filler particles create for the macromolecules of the polymer matrix. It is due to these effects that the glass transition temperature decreases.

In addition to glass process, the conditions of preparation of polymer composites filled with organoclay significantly affect melting processes. The shape of the maxima on Fig. 3, which correspond to the melting of the polymer matrix, changes significantly with an increase in the processing time of the system using an extruder. Already after 2 min of mixing, an additional peak appears on the graphs after the main maximum, the intensity of which increases with increasing processing time. The nature of the additional maximum is probably related to the segregation of a separate crystalline phase. In our opinion, this phase is formed from the polymer-filler boundary layer, due to which crystalline formations are formed, the melting of which requires more energy.

Fig. 4 also presents the dependence of the melting temperature (T_m) of the PEO crystalline phase on the extrusion time. It is observed that with mechanical mixing without extrusion, the melting temperature of the system is almost equal to the T_m of pure PEO. This behavior of T_m indicates the absence of influence of non-intercalated MMT on the melting processes of the polymer matrix. As the mixing time increases, the melting point drops sharply, which correlates with a decrease in the glass transition temperature. When processed for 2 minutes, the melting point decreased by 13 ºC.

From the thermophysical data, the degree of crystallinity of composites based on PEO was calculated using equation (3):

$$\chi_c = \frac{\Delta H_m}{\Delta H_{m,c}} , \qquad (3)$$

where ΔH_m is measured enthalpy of melting, $\Delta H_{m,c}$ is melting enthalpy of 100 % crystalline polymer (for PEO, = 165.5 J/g).

Fig. 4 shows the dependence of the degree of crystallinity on the extrusion time. It is observed that the degree of crystallinity depends on the time of polymer composites mixing. It should be noted that the behavior of the degree of crystallinity determined from DSC data is similar to the behavior of the degree of crystallinity obtained from the results of X-ray structural analysis. Also, this behavior is fully correlated with the behavior of the melting and glass transition temperatures and is explained by the processes of intercalation of polymer macromolecules into the interlayer space of NC.

C. Electrical Characteristics of the PEO-NC System.

The structural organization of nanoclay in the polymer matrix significantly affects the electrical properties of the system. Fig. 5a shows the frequency dependences of electrical conductivity for polymer composites with different preparation times. Frequency dependences for unfilled PEO and composites based on it are non-linear. They have a plateau when electrical conductivity does not depend on frequency and a sharp rise in the high-frequency region. This behavior of dependencies is typical for most polymers [21-23]. Values of electrical conductivity that frequency independent (plateau region) correspond to electrical conductivity at direct current.

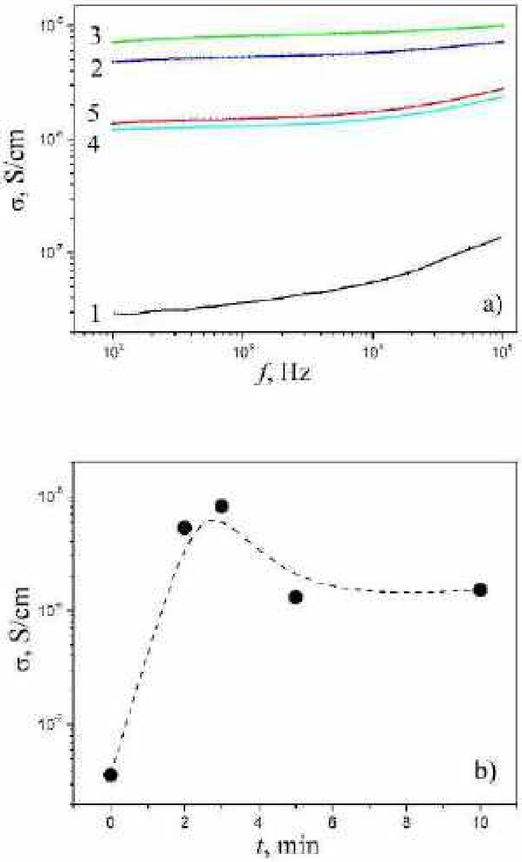


Fig. 5. Frequency dependences of electrical conductivity (a) and dependence of electrical conductivity on preparation time (b) for polymer composites based on PEO and NC, produced by extrusion during: 1 – 0 min; 2 – 2 min; 3 – 3 min; 4 – 5 min; 5 – 10 min.

Fig. 5b shows the dependence of electrical conductivity at direct current on the preparation time. Electrical conductivity in fig. 5b behaves non-linearly. It first increases, reaching a maximum after 2-3 minutes of processing, and then almost does not change. Similar extreme behavior was observed for thermophysical characteristics. It is explained by intercalation processes, which reach their maximum at 3 min of mixing.

CONCLUSIONS

The paper investigates the effect of mixing time using a composite extruder based on PEO-1000 and organomodified nanoclay on its structural, thermophysical and electrical characteristics. It was found that the mixing time significantly affects both the structural and functional characteristics of the studied systems. According to the results of the analysis of X-ray scattering data, it was established that the interlayer distance of NC increases with increasing processing time and reaches its maximum value at 3-5 minutes of mixing. At the same time, the maximum intercalation of PEO macromolecules in the interlayer space of NC takes place. That is why extreme values of structural (degree of crystallinity), thermophysical (melting and glass transition temperatures) and electrical (direct current electrical conductivity) characteristics of the investigated composites are observed at this value of extrusion time. With intercalation and partial delamination of NC, its surface area, which is able to interact with PEO macromolecules, increases. The developed surface of the nanofiller blocks the processes of free movement of macromolecules and creates obstacles for their stacking in crystal structures. Due to this process, the degree of crystallinity of the investigated composites significantly decreases with increasing processing time.

Therefore, the mixing time of the components of the polymer composite using an extruder equal to 3 minutes is optimal. At the same time, the mechanical energy absorbed by the system is sufficient for the maximum delamination of NC tactoids. A further increase in the processing time does not lead to an improvement in the structural and functional characteristics of the studied systems.

Acknowledgment

This study was funded from the State Budget of Ukraine within the framework of the scientific project No. 0121U100658 “Creation of new multifunctional nanocomposite polymer materials containing carbon nanotubes” commissioned by the Ministry of Education and Science of Ukraine.

References

[1] M. Muhammed Shameem, S.M. Sasikanth, R. Annamalai and R. Ganapathi Raman, “A brief review on polymer nanocomposites and its applications,” Materials Today: Proceedings, vol. 45 (2), pp. 2536–2539, 2021.

[2] M. Khodakarami and M. Bagheri, “Recent advances in synthesis and application of polymer nanocomposites for water and wastewater treatment,” Journal of Cleaner Production, vol. 296. p. 126404, 2021.

[3] Z. Feng, K.H. Adolfsson, Y. Xu, H. Fang, M. Hakkarainen and M. Wu, “Carbon dot/polymer nanocomposites: From green synthesis to energy, environmental and biomedical applications,” Sustainable materials and technologies, Vol. 29. p. e00304, 2021.

[4] T.S. Daitx, L.N. Carli, J.S. Crespo and R.S. Mauler, “Effects of the organic modification of different clay minerals and their application in biodegradable polymer nanocomposites of PHBV,” Applied Clay Science, vol. 115. pp. 157–164, 2015.

[5] L.F.B.L. Pontes, J.E.G. de Souza, A. Galembeck and C.P. de Melo, “Gas sensor based on montmorillonite/polypyrrole composites prepared by in situ polymerization in aqueous medium,” Sensors and Actuators B: Chemical, vol. 177, pp. 1115–1121, 2013.

[6] Y. Zhao and Y. Wang, “Tailored Solid Polymer Electrolytes by Montmorillonite with High Ionic Conductivity for Lithium-Ion Batteries,” Nanoscale Research Letters, vol. 14, p. 366, 2019.

[7] J. Hu, X. Zhao, J. Xie, Y. Liu, S. Sun, “Effect of organic Na+-montmorillonite on the dielectric and energy storage properties of polypropylene nanocomposites with polypropylene-graft-maleic anhydride as compatibilizer,” Journal of Applied Polymer Science, vol. 139, Is.17, p. 52047, 2022.

[8] A. Amari, F. Mohammed Alzahrani, K. Mohammedsaleh Katubi, N. Salem Alsaiari, M.A. Tahoon and F. Ben Rebah, “Clay-Polymer Nanocomposites: Preparations and Utilization for Pollutants Removal,” Materials (Basel), vol. 14 (6), p. 1365, 2021.

[9] H.V. Torrecillas, L.C. Costa and A.M.C. Souza, “Influence of mixing protocol on the morphology and mechanical properties of PP/SEBS/MMT and PP/SEBS/PPgMA/MMT blends,” Polymer Testing, vol. 72, pp. 322–329, 2018.

[10] M.R. Kamal, J.U. Calderon and B.R. Lennox, “Surface Energy of Modified Nanoclays and Its Effect on Polymer/Clay Nanocomposites,” Journal of Adhesion Science and Technology, vol. 23 (5), pp. 663–688, 2009.

[11] P. Kumar, K.P. Sandeep, S. Alavi, V.D. Truong and R.E. Gorga, “Preparation and characterization of bio-nanocomposite films based on soy protein isolate and montmorillonite using melt extrusion,” Journal of Food Engineering. vol. 100, no 3, pp. 480–489, 2010.

[12] E.A. Lysenkov, S.A. Bilyi, V.V. Klepko and L.P. Klymenko, “Features of Intercalation Processes in Polymer Nanocomposites Based on Oligoethylene Glycol and Organoclay,” IEEE 11th International Conference Nanomaterials: Applications & Properties (NAP), 2021. DOI: 10.1109/NAP51885.2021.9568615

[13] E. Lysenkov, V. Klepko, L. Bulavin and N. Lebovka, “Physico-Chemical Properties of Laponite®/Polyethylene-oxide Based Composites,” Chemical Record, p. e202300166, 2023.

[14] L. Delva, K. Ragaert, J. Degrieck and L. Cardon, “The Effect of Multiple Extrusions on the Properties of Montmorillonite Filled Polypropylene,” Polymers, vol. 6, pp. 2912–2927, 2014.

[15] E.A. Lysenkov, V. Klepko and M.M. Lazarenko, “Structure-Properties Relationships of Nanocomposites Based on Polyethylene Oxide and Anisometric Nanoparticles,” in Nanomaterials and Nanocomposites, Nanostructure Surfaces and Their Applications, O. Fesenko and L. Yatsenko, Eds. Switzerland: Springer International Publishing, vol. 279, pp. 409-437, 2023.

[16] E. Lysenkov and L. Klymenko, “Determination of the effect of carbon nanotubes on the microstructure and functional properties of polycarbonate-based polymer nanocomposite materials,” Eastern European Journal of Enterprise Technologies, vol. 12, no 4, pp. 53–60, 2021.

[17] N. Choi, Y. Son, T.-H. Kim, Y. Park and Y. Hwang, “Adsorption behaviors of modified clays prepared with structurally different surfactants for anionic dyes removal,” Environmental Engineering Research, vol. 28 (2), p. 210076, 2023.

[18] F. Guo, S. Aryana, Y. Han and Y. Jiao, “A Review of the Synthesis and Applications of Polymer–Nanoclay Composites,” Appl. Sci., vol. 8, p. 1696, 2018.

[19] C.G. Pope, “X-Ray Diffraction and the Bragg Equation,” J. Chem. Educ., vol. 74, no 1, p. 129, 1997.

[20] E. Aytunga Arik Kibar and Ferhunde Us, “Evaluation of Structural Properties of Cellulose Ether-Corn Starch Based Biodegradable Films,” International Journal of Polymeric Materials and Polymeric Biomaterials, vol. 63, pp. 342–351, 2014.

[21] A. Buketov, S. Smetankin, E. Lysenkov, K. Yurenin, O. Akimov, S. Yakushchenko and I. Lysenkova, “Electrophysical Properties of Epoxy Composite Materials Filled with Carbon Black Nanopowder,” Advances in Materials Science and Engineering, vol. 2020, Art. ID 6361485, 2020.

[22] X. Xia, Y. Wang, Z. Zhong and G.J. Weng, “A frequency-dependent theory of electrical conductivity and dielectric permittivity for graphene-polymer nanocomposites,” Carbon, vol. 111, pp. 221–230, 2017.

[23] C. Tsonos, “Comments on frequency dependent AC conductivity in polymeric materials at low frequency regime,” Current Applied Physics, vol. 19, Is. 4, pp. 491–497, 2019.

Structural Features of Porous-GaAs and its Potential in Heterostructural Buffer Layers

Yana Suchikova
The Department of Physics and Methods of Teaching Physics
Berdyansk State Pedagogical University
Berdyansk, Ukraine
yanasuchikova@gmail.com

Sergii Kovachov
The Department of Physics and Methods of Teaching Physics
Berdyansk State Pedagogical University
Berdyansk, Ukraine
essfero@gmail.com

Zhakyp Karipbaev
L.N. Gumilyov Eurasian National University
2 Satpayev Str.
Nur-Sultan, 010008, Kazakhstan
karipbayev_zht_1@enu.kz

Ihor Bohdanov
The Department of Physics and Methods of Teaching Physics
Berdyansk State Pedagogical University,
Berdyansk, Ukraine
naukabdpu@gmail.com

Anastasiia Lysak
Berdyansk State Pedagogical University, Berdyansk, Ukraine
Institute of Physics, Polish Academy of Sciences,
Warsaw, Poland
https://orcid.org/0000-0002-3114-6526

Anatoli I. Popov
Institute of Solid State Physics
University of Latvia
8 Kengaraga str.
LV-1063, Riga, Latvia
popov@latnet.lv

***Abstract* — This study investigates porous gallium arsenide's structural and morphological properties. Based on EDX analysis data, it is established that Ga and As dominate the material composition, indicating its consistency with gallium arsenide, while there is a limited presence of oxygen. SEM analysis revealed a uniform distribution of pores on the surface, forming intricate tracks and chains. Although por-GaAs possess a porous structure, their crystalline properties, as demonstrated through XRD-spectroscopy, remain preserved and are similar to monocrystalline GaAs. The spectroscopic analysis detected vibrational processes characteristic of GaAs and features associated with the porosity and structure of por-GaAs.**

Keywords* — *Gallium arsenide, por-GaAs, electrochemical etching, morphology, Raman investigation, XRD-spectroscopy

I. Introduction

Modern technologies' continuous development and evolution demand the creation of new and improved materials. Nanostructured materials, in particular, have garnered significant interest as they offer superior characteristics to traditional materials. Such materials pave the way for developing new devices, instruments, and systems. However, to attain the optimal characteristics of these materials, a deep understanding of their structure and properties at the micro and nano levels is essential.

Gallium compounds hold a special significance in nanostructured materials due to their high potential in electronics and optoelectronics [6-9]. Due to their superior electron mobility, which surpasses silicon by a factor of five, GaAs has long been recognized as a promising material for developing advanced electronic and optoelectronic devices [10]. However, the characteristics of GaAs are complicated by surface phenomena and defects, leading to high surface recombination rates [12, 13]. In this regard, research efforts are directed towards finding optimal processing and passivation methods for the GaAs surface to enhance its properties and extend its longevity [14, 15].

Recently, porous GaAs have established its footing in scientific research due to its unique properties, positioning it as an alternative to traditional GaAs in applications where photonic and optoelectronic attributes are crucial [16, 17]. The structural features of porous GaAs open new avenues for controlling electron and phonon interactions within the material [18, 19]. Additionally, porous layers are a reliable buffer layer for heterostructure creation [20]. They act as a "soft" substrate, alleviating excess stress that may arise between the template and the grown structure due to lattice mismatch [21, 22]. In this regard, maintaining the crystallinity of the porous layer is essential, which poses a significant technological challenge, as pore formation on the monocrystal surface can distort its lattice.

This paper presents a straightforward synthesis method for porous GaAs layer on mono-GaAs' surface and delves into its morphological, chemical, structural, and phononic characteristics.

II. Experiment

A. Sample preparation

Monocrystalline samples of gallium arsenide, grown by the Czochralski method, were used in the experiment. We used cubic n-type mono-GaAs, doped with Sb to a charge carrier concentration of 2.3x10^{18} cm^{-3}. The samples' orientation was set in the (111) direction.

Sample preparation involved chemical etching in a hydrochloric acid solution to achieve a smooth surface and remove the oxide layer. The next step was cleaning the samples with vinegar and ethyl alcohol. It is essential to note that the chemical etching was performed before the experiment to prevent the potential re-oxidation of the semiconductor material.

B. Experimental methodology

Porous gallium arsenide (por-GaAs) was formed on mono-GaAs' surface using an anodic electrochemical etching technique. A standard three-electrode electrochemical cell was employed. The system consisted of a working electrode made of monocrystalline gallium arsenide, a reference electrode, and a platinum counter electrode. Electrochemical etching was performed in an electrolyte prepared using hydrochloric acid (HCl) and nitric acid (HNO$_3$) in distilled water. The solvent concentration in the electrolyte was chosen in the following ratio: 2 M for HCl and 1.5 M for HNO$_3$.

The experiment was conducted in a potentiostatic mode. Samples were etched in the electrolyte solution for 5 minutes. Observations during the etching process for current density

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

showed a maximum value of j=200 mA/cm on the 4th minute of etching. After this, the current began to decrease gradually, indicating the completion of the active pore formation process. Etching was stopped, and samples were removed from the electrolyte solution, rinsed in hydrogen peroxide, and dried in a stream of atomic nitrogen to stabilize surface properties.

Thus, the simplest version of electrochemical etching was used. The peculiarity of the process was only the use of specific electrolyte composition and fairly aggressive short-term etching conditions. The hypothesis was to obtain a densely packed porous layer with crystallographically oriented pores to preserve the crystalline characteristics of bulk-GaAs. It was anticipated that using an accurate potentiostatic mode and optimized dissolution conditions would allow obtaining a structured gallium arsenide surface with high homogeneity and controlled porosity.

C. Sample Characterization

Scanning electron microscopy was used on an SEM device for surface morphology studies. Additionally, for elemental composition analysis, energy-dispersive X-ray spectroscopy (EDX) was used.

Raman spectroscopy of samples was conducted using a RENISHAW inVia Reflex microscope. Samples were irradiated with a 532 nm wavelength laser using a 2400 nm grating. Spectra were recorded in the range of 100-1000 cm^{-1}. Each measurement lasted 10 seconds, and five accumulations were carried out for each sample to ensure the highest data accuracy.

Analysis of the sample structure using X-ray diffractometry was conducted on the XRD Drone-3M device. Measurements were made in the range of 2θ angles from 10° to 80° with a step of 0.01 degrees, allowing detailed information about the material's crystalline structure to be obtained.

III. Results

A. EDX Analysis of Porous Gallium Arsenide

EDX (Energy Dispersive X-ray Spectroscopy) analysis was conducted to characterize the elemental composition of porous gallium arsenide. The results of the analysis are presented in the form of an EDX spectrum (Figure 1) and EDX mapping (Figure 2).

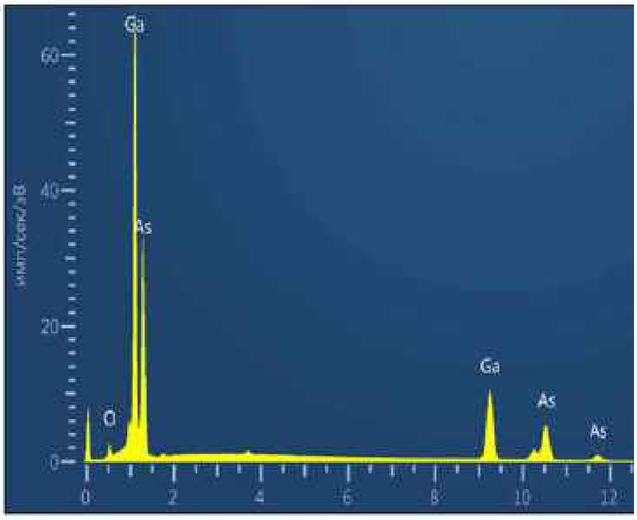


Fig. 1. EDX spectrum of porous gallium arsenide.

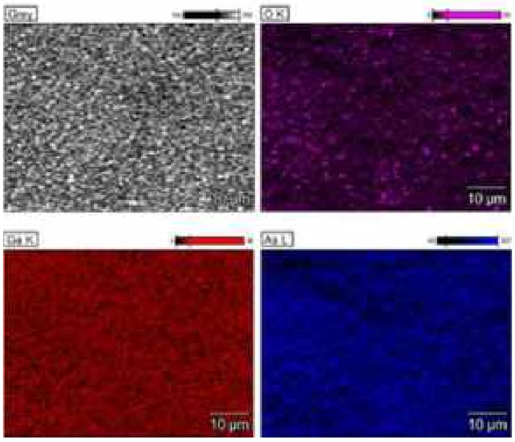


Fig. 2. EDX mapping of porous gallium arsenide.

Peaks corresponding to Ga, As, and O are observed in the EDX spectrum (Figure 1), reflecting the primary structure of porous gallium arsenide. The mass and atomic percentages of these elements are provided in Table 1.

TABLE I. Mass And Atomic Percentages Of Elements O, Ga, And As In Porous Gallium Arsenide (Por-GaAs) Based On The Edx Analysis

Element	Compound	
	At. %	Wt. %
O-K	2.13	0.48
Ga-L	42.41	41.38
As-L	55.46	58.14

Based on the spectroscopy data, Ga and As constitute most of the material, aligning with the gallium arsenide structure. Oxygen is present in considerably smaller amounts, indicating limited oxidation of the surface of porous GaAs following the etching process.

EDX mapping (Figure 2) illustrates the distribution of these elements on the por-GaAs surface. The mapping data reveal that Ga and As completely cover the surface, forming a homogeneous structure. Oxygen, conversely, is present in the form of dispersed inclusions throughout the surface, attesting to its limited presence.

According to the EDX analysis data, the atomic percentage ratio of Ga to As is approximately 1:1.31. This deviates slightly from the stoichiometric ratio of 1:1 that is expected for monocrystalline GaAs. This discrepancy can be attributed to several factors. Firstly, there might be a depletion of Ga atoms in porous GaAs during the electrochemical etching process. Secondly, there may be oxidation of Ga on the por-GaAs surface, leading to the formation of a small amount of Ga-O.

Despite these deviations, the Ga to As ratio remains close to the balance for monocrystalline GaAs, indicating the effectiveness of the etching process.

B. SEM Analysis

Figure 3 displays the SEM image of the porous GaAs surface. The pores exhibit the "111" crystallographic orientation. However, unlike traditional "Crysto pores," where pores intersect at certain angles, in this instance, the pores predominantly have a parallel orientation to each other.

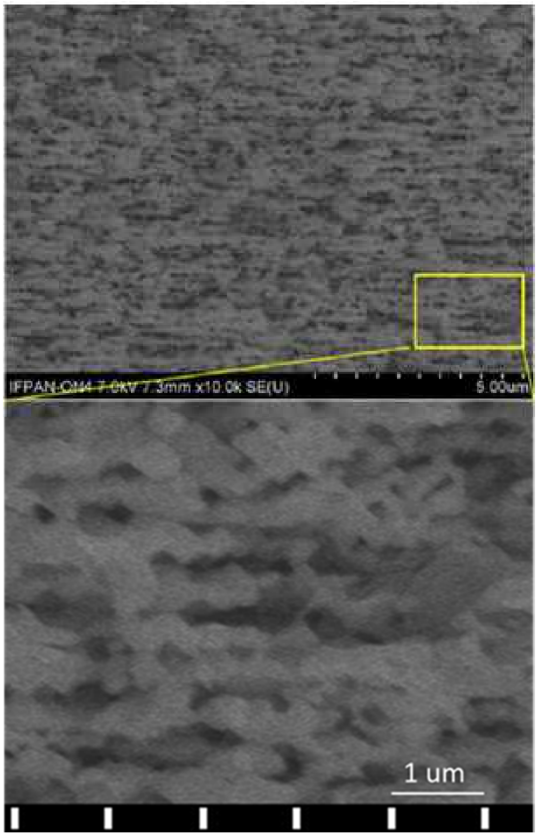


Fig. 3. - SEM image of por-GaAs surface.

The pores form channels, tracks, and chains, which occasionally intersect with each other. Oxidation islands are not observed on the por-GaAs surface, even though EDX analysis data revealed the presence of trace amounts of oxygen. This may suggest exceptionally localized oxidation.

Inter-pore spaces construct a bead-like structure. The pores evenly cover the surface, forming an intricate macromorphological landscape. The formed tracks range from 1 to 3 µm, although ultra-long tracks exceeding 5 µm are also observed. They primarily have a cross-sectional size ranging from 100 to 200 nm, with aspect ratios ranging from 20 to 50. Each track exhibits an uneven structure along its length, creating a "bead" effect.

Figure 4 provides a cross-sectional view of the formed porous structure. The thickness of the porous layer demonstrates heterogeneity. The thickness of the porous layer reaches a maximum value of 35 µm, whereas the minimum value is around 20 µm. The pores don't exhibit a distinct growth direction into the crystal's thickness.

C. XRD Analysis

XRD spectra of por-GaAs and the theoretically calculated pattern for mono-GaAs (sourced from Crystallography Open Database, COD ID 9008845) are displayed in Figure 5. Intense peaks are observed at 2θ = 27.3, 31.6, 53.73, 66.0°, corresponding to reflections from planes (111), (220), (311), and (400), respectively. The spectrum for por-GaAs exhibits an excellent match with its monocrystalline counterpart, concluding that no lattice reconstruction occurred. A comparison of the main parameters of the crystalline structure of por-GaAs and mono-GaAs is presented in Table 2.

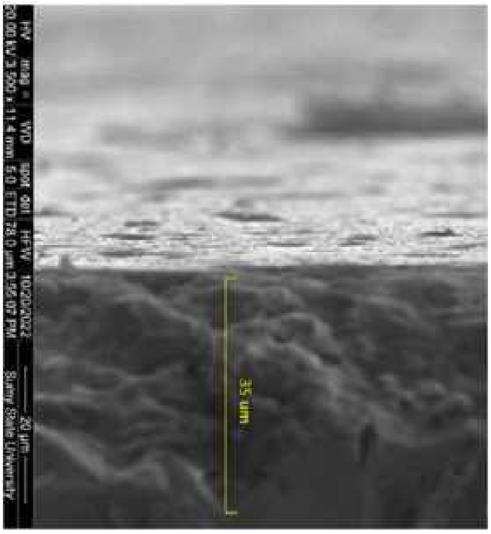

HV mag WD spot det HFW 10/20/2022
20.00 kV 3 500 x 11.4 mm 5.0 ETD 28.0 µm 3:56:07 PM
20 µm
Sumy State University
35 um

Fig. 4. - Cross-sectional view of porous-GaAs.

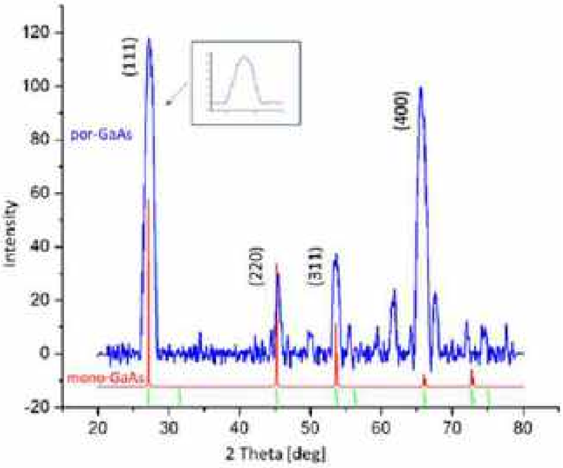


Fig. 5. - Cross-sectional view of porous-GaAs.

TABLE II. Structural Parameters Of Por-GaAs (Experimental) And Mono-GaAs (Theoretical)

Parameter	Value	
	por-GaAs	**mono-GaAs**
Crystal system	Cubic	Cubic
Lattice type	F	F
Space group name	F-43m	Fm-3m
Lattice parameters	a= 5.654 Å	a= 5.654 Å
Unit-cell volume	180.73 Å^3	180.72Å^3

The peaks are sufficiently narrow and sharp, indicating a well-crystallized structure and the absence of additional phases. The peak at 2θ = 31.6° exhibits a barely noticeable shift to the right, which could be a manifestation of quantum size effects. Additionally, a considerable increase in the peak intensity at 2θ = 66.0° is observed compared to the theoretical value of the monocrystalline counterpart. This could be attributed to increased crystal orientation along the (400) plane due to the pore formation process.

Overall, the examination of the XRD spectra indicates that despite substantial macroscopic restructuring resulting in a porous structure, the atomic-level structure of gallium arsenide remains stable and monocrystalline. This emphasizes the high stability of the GaAs crystal lattice and suggests the potential of using por-GaAs/mono-GaAs as a material with novel properties based on the combination of monocrystalline and porous morphology.

To determine the average crystallite size, the Scherrer equation was applied:

$$D = \frac{K\lambda}{\beta \cos\theta}, \tag{1}$$

where: D is the average crystallite size; K represents the crystallite shape factor; λ is the wavelength of the CuKα1 X-ray radiation (1.540598 Å); B is the full width at half maximum (FWHM) in radians; θ is the diffraction angle (within Bragg's limits).

The shape factor K for the crystallite was taken as K = 0.89 due to the cubic crystalline symmetry [23]. The calculation was carried out for the most intense peak, corresponding to reflection from the (111) plane. The calculation of the average crystallite size yielded a value of D = 60.49 nm.

D. Raman Analysis

Figure 6 displays the first-order combination scattering spectrum, recorded at room temperature under non-resonant conditions for synthesized porous GaAs. The spectrum predominantly showcases two intense peaks situated around 266 and 287 cm^{-1}. These modes display minor rightward shifts compared to bulk-GaAs frequencies (269 and 292 cm^{-1} for TO and LO, respectively [24, 25]). Rightward shifts in frequencies for porous GaAs (compared to bulk GaAs) might be attributed to structural peculiarities of the porous material and additional surface defects influencing phonon properties. The strong mode at 266 cm^{-1} reflects the dominant vibrational processes of atoms in the GaAs lattice. The peak at 287 cm^{-1} of medium intensity may correspond to vibrations typical for bulk GaAs. Its presence confirms that the primary material properties are retained despite the porous structure.

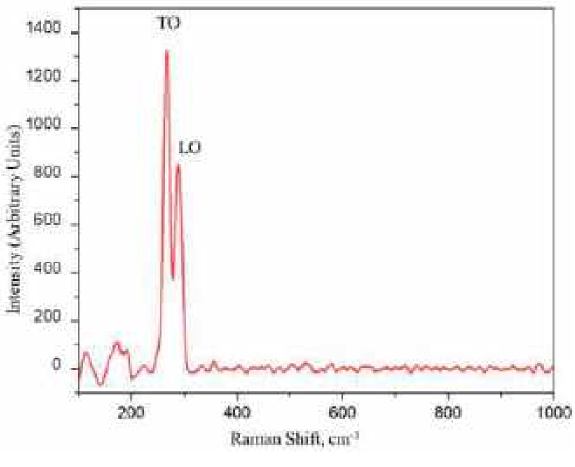


Fig. 6. - Raman spectrum por-GaAs.

The peak observed at 168 cm^{-1} can be attributed to As–As vibrations or might be influenced by interactions between Ga and As atoms [26]. The weak intensity of the peak at 111 cm^{-1} indicates low-energy vibrational processes. In theoretical studies of bulk GaAs, this mode is rarely encountered, suggesting its presence might indicate the peculiarities of the porous structure or defects in the lattice.

IV. Discussion

Based on our observations, our electrochemical etching method, with specifically tailored conditions and electrolyte composition, is effective for producing porous gallium arsenide with high uniformity and controllable porosity.

These findings validate our hypothesis about the feasibility of producing a densely packed porous layer with crystallographically oriented pores that retain the crystalline properties of bulk GaAs. Consequently, our experimental results unveil vast opportunities for further research and applications in areas like heterostructures, photonics, optoelectronics, and other scientific and technological fields where materials' structural and electronic properties are crucial.

Porous gallium arsenide (por-GaAs) is particularly interesting due to its unique structural and morphological characteristics and potential application in developing heterostructures. In creating heterostructures, por-GaAs can serve as a buffer layer, facilitating harmonious integration of different materials or enhancing the growth quality of subsequent layers.

One of the critical advantages of porous gallium arsenide is its retention of crystalline properties similar to that of bulk GaAs. Preserving crystallinity is crucial for heterostructures as it ensures an excellent electronic and optical match between the different layers of the structure. This, in turn, can enhance the performance and quality of devices based on such heterostructures.

Furthermore, porous gallium arsenide offers several advantages over its bulk counterpart. Its porous structure allows for enhanced adhesion between different heterostructure layers, which can mitigate internal stresses and boost thermal stability. Moreover, porosity can act as a mechanism for trapping defects and other unwanted impurities, thereby enhancing the quality of the heterostructure.

Thus, considering the aforementioned characteristics, porous gallium arsenide may find extensive application in traditional fields like photonics or optoelectronics and in developing innovative heterostructures for next-generation devices.

Conclusions

In the scope of this study, we have demonstrated the methodology and peculiarities of synthesizing porous gallium arsenide (por-GaAs) based on anodic electrochemical etching in an electrolyte solution based on HCl and HNO_3, utilizing high potential values and short processing time. The conclusions of our research are presented below:

- The efficacy of the anodic electrochemical etching method for forming porous gallium arsenide (por-GaAs) on the surface of monocrystalline gallium arsenide (mono-GaAs) was successfully investigated and confirmed.
- SEM analysis data validate the formation of evenly distributed pores on the surface, manifesting as tracks and chains. Employing a defined electrolyte composition and aggressive short-term etching conditions facilitated the acquisition of a densely packed porous layer with crystallographically oriented pores.

- Despite por-GaAs possessing a porous structure, their crystalline properties, as per XRD spectroscopy, remain comparable with monocrystalline GaAs. This emphasizes the chosen method's effectiveness in retaining the material's crystallographic characteristics.
- Spectroscopic analysis not only indicates primary vibrational processes typical for GaAs but also reveals additional features that reflect the porosity and structural peculiarities of por-GaAs.

The results affirm the potential application of por-GaAs as a prospective material for heterostructure buffer layers, potentially heralding new horizons for developing innovative devices and systems.

Acknowledgment

The study was supported by the Ministry of Education and Science of Ukraine via Project No. 0122U000129 “The search for optimal conditions for nanostructure synthesis on the surface of A3B5, A2B6 semiconductors and silicon for photonics and solar energy” and Project No. 0121U10942 “Theoretical and methodological bases of system fundamentalization of the future nanomaterials experts training for productive professional activity”. In addition, the research was partly supported by COST Action CA20129 “Multiscale Irradiation and Chemistry Driven Processes and Related Technologies” (MultIChem), and COST Action CA20126 - Network for research, innovation, and product development on porous semiconductors and oxides (NETPORE). A.P. thanks to the Institute of Solid-State Physics, University of Latvia. ISSP UL as the Center of Excellence is supported through the Framework Program for European universities, Union Horizon 2020, H2020-WIDESPREAD-01–2016–2017-TeamingPhase2, under Grant Agreement No. 739508, CAMART2 project.

References

[1] H. Klym, I. Karbovnyk, M.C. Guidi, O. Hotra, and A.I. Popov, "Optical and Vibrational Spectra of CsCl-Enriched GeS_2-Ga_2S_3 Glasses," Nanoscale Research Letters, vol. 11, no. 1, art. no. 132, 2016.

[2] V.P. Savchyn, A.I. Popov, O.I. Aksimentyeva, H. Klym, Y.Y. Horbenko, V. Serga, A. Moskina, and I. Karbovnyk, "Cathodoluminescence characterization of polystyrene-$BaZrO_3$ hybrid composites," Low Temperature Physics, vol. 42, no. 7, pp. 760-763, 2016.

[3] S. Yana, "Porous indium phosphide: Preparation and properties," in Handbook of Nanoelectrochemistry: Electrochemical Synthesis Methods, Properties, and Characterization Techniques, 2016, pp. 283–306.

[4] Y. Suchikova, S. Kovachov, I. Bohdanov, ..., A. Moskina, and A. Popov, "Characterization of $Cd_xTe_yO_z$/CdS/ZnO Heterostructures Synthesized by the SILAR Method," Coatings, vol. 13, no. 3, pp. 639, 2023.

[5] H. Klym, I. Karbovnyk, S. Piskunov, and A.I. Popov, "Positron annihilation lifetime spectroscopy insight on free volume conversion of nanostructured $MgAl_2O_4$ ceramics," Nanomaterials, vol. 11, no. 12, pp. 3373, 2021.

[6] H. Klym, A. Ingram, O. Shpotyuk, O. Hotra, and A.I. Popov, "Positron trapping defects in free-volume investigation of Ge–Ga–S–CsCl glasses," Radiation Measurements, vol. 90, pp. 117-121, 2016.

[7] H. Klym, I. Karbovnyk, A. Luchechko, Y. Kostiv, V. Pankratova, and A.I. Popov, "Evolution of free volumes in polycrystalline $BaGa_2O_4$ ceramics doped with Eu3+ ions," Crystals, vol. 11, no. 12, pp. 1515, 2021.

[8] A. Usseinov, Z. Koishybayeva, A. Platonenko, ..., Y. Suchikova, and A.I. Popov, "Ab-Initio Calculations of Oxygen Vacancy in Ga_2O_3 Crystals," Latvian Journal of Physics and Technical Sciences, vol. 58, no. 2, pp. 3–10, 2021.

[9] Z.T. Karipbayev, K. Kumarbekov, I. Manika, ..., Y. Suchikova, and A.I. Popov, "Optical, Structural, and Mechanical Properties of $Gd_3Ga_5O_{12}$ Single Crystals Irradiated with 84Kr+ Ions," Physica Status Solidi (B) Basic Research, vol. 259, no. 8, pp. 2100415, 2022.

[10] R. Trommer and M. Cardona, "Resonant Raman scattering in GaAs," Physical Review B, vol. 17, no. 4, pp. 1865, 1978.

[11] Y. B. Bolkhovityanov and O. P. Pchelyakov, "GaAs epitaxy on Si substrates: modern status of research and engineering," Physics-Uspekhi, vol. 51, no. 5, pp. 437, 2008.

[12] S.O. Vambol, I.T. Bohdanov, V.V. Vambol, ..., T.P. Nestorenko, and S.V. Onyschenko, "Formation of filamentary structures of oxide on the surface of monocrystalline gallium arsenide," Journal of Nano- and Electronic Physics, vol. 9, no. 6, pp. 06016, 2017.

[13] F. Z. Elamri, F. Falyouni, A. Kerkour-El Miad, and D. Bria, "Effect of defect layer on the creation of electronic states in GaAs/GaAlAs multi-quantum wells," Applied Physics A, vol. 125, pp. 1-12, 2019.

[14] Y. Suohikova, S. Vambol, V. Vambol, N. Mozaffari, and N. Mozaffari, "Justification of the most rational method for the nanostructures synthesis on the semiconductors surface," Journal of Achievements in Materials and Manufacturing Engineering, vol. 92, no. 1-2, 2019.

[15] S. Manna, H. Huang, S. F. C. da Silva, C. Schimpf, M. B. Rota, B. Lehner, ... and A. Rastelli, "Surface passivation and oxide encapsulation to improve optical properties of a single GaAs quantum dot close to the surface," Applied Surface Science, vol. 532, art. no. 147360, 2020.

[16] A. P. Oksanich, S. E. Pritchin, M. G. Kogdas, A. G. Kholod, and M. G. Dernova, "Pd/Porous GaAs in the Manufacture of Schottky Diodes," in 2019 IEEE International Conference on Modern Electrical and Energy Systems (MEES), September 2019, pp. 110-113.

[17] A. Hernández, Y. Kudriavtsev, C. Salinas-Fuentes, C. Hernández-Gutierrez, and R. Asomoza, "Optical properties of porous GaAs formed by low energy ion implantation," Vacuum, vol. 171, art. no. 108976, 2020.

[18] Y. Suchikova, S. Kovachov, and I. Bohdanov, "Formation of oxide crystallites on the porous GaAs surface by electrochemical deposition," Nanomaterials and Nanotechnology, vol. 12, 2022.

[19] A. P. Oksanich, S. E. Pritchin, M. G. Kogdas, A. G. Kholod, and M. G. Dernova, "Pd/Porous GaAs in the manufacture of Schottky diodes," in 2019 IEEE International Conference on Modern Electrical and Energy Systems (MEES), September 2019, pp. 110-113.

[20] Y. Suchikova, S. Kovachov, I. Bohdanov, ..., V. Pankratov, and A.I. Popov, "Study of the structural and morphological characteristics of the $Cd_xTe_yO_z$ nanocomposite obtained on the surface of the CdS/ZnO heterostructure by the SILAR method," Applied Physics A: Materials Science and Processing, vol. 129, no. 7, pp. 499, 2023.

[21] L. Beji, B. Ismaıl, L. Sfaxi, F. Hassen, H. Maaref, and H. B. Ouada, "Critical layer thickness enhancement of InAs overgrowth on porous GaAs," Journal of Crystal Growth, vol. 258, no. 1-2, pp. 84-88, 2003.

[22] Y. Suchikova, S. Kovachov, A. Lazarenko, and I. Bohdanov, "Research of synthesis conditions and structural features of heterostructure $Al_XGa_{1-X}As$/GaAs of the ‘desert rose’ type," Applied Surface Science Advances, vol. 12, p. 100327, 2022.

[23] J. Langford and A. Wilson, "Scherrer after sixty years: A survey and some new results in the determination of crystallite size," J. Appl. Crystallogr., vol. 11, pp. 102–113, 1978

[24] M. T. Constant, A. Bellarbi, A. Lorriaux, and B. Grimbert, "Raman scattering characterization of processing effects on GaAs planar photoconductors," in Spectroscopic Characterization Techniques for Semiconductor Technology IV, vol. 1678, pp. 137-146, July 1992

[25] B. Prévot and J. Wagner, "Raman characterization of semiconducting materials and related structures," Progress in Crystal Growth and Characterization of Materials, vol. 22, no. 4, pp. 245-319, 1991.

[26] S. V. Sorokin, P. S. Avdienko, I. V. Sedova, D. A. Kirilenko, V. Y. Davydov, O. S. Komkov, and S. V. Ivanov, "Molecular beam epitaxy of layered group III metal chalcogenides on GaAs (001) substrates," Materials, vol. 13, no. 16, p. 3447, 2020.

Optimization of $Cd_xTe_yO_z$ Synthesis Modes by the SILAR Method

Yana Suchikova
The Department of Physics and Methods of Teaching Physics
Berdyansk State Pedagogical University
Berdyansk, Ukraine
yanasuchikova@gmail.com

Sergii Kovachov
The Department of Physics and Methods of Teaching Physics
Berdyansk State Pedagogical University
Berdyansk, Ukraine
essfero@gmail.com

Zhakyp Karipbaev
L.N. Gumilyov Eurasian National University
2 Satpayev Str.
Nur-Sultan, 010008, Kazakhstan
karipbayev_zht_1@enu.kz

Yaroslav Zhydachevskyy
Berdyansk State Pedagogical University Berdyansk, Ukraine
Institute of Physics, Polish Academy of Sciences,
Poland
zhydach@ifpan.edu.pl899

Ihor Bohdanov
Berdyansk State Pedagogical University
Berdyansk, Ukraine
naukabdpu@gmail.com

Anatoli I. Popov
Institute of Solid State Physics University of Latvia
8 Kengaraga str.
LV-1063, Riga, Latvia
popov@latnet.lv

***Abstract* — This study investigates the synthesis of $Cd_xTe_yO_z$ films grown via the SILAR method on the CdS/ZnO surface. It was confirmed that the number and duration of treatment cycles influence the structural and phase properties of the films. The results indicate the potential of the SILAR method for developing nanocomposite materials with specific characteristics.**

***Keywords* — SILAR method, heterostructures, nano-composites, films, oxides**

I. Introduction

Nanocomposite films have attracted researchers' attention due to their unique properties and potential in various applications, such as photovoltaics, photocatalysis, and optical devices [1, 2]. Special attention is paid to materials with nanometer-sized structural elements [3, 4].

From a modern nanotechnology and materials science perspective, oxide semiconductors, heterostructures, and composites with an inherent oxide layer on the surface emerge as promising and intriguing research subjects [5, 6]. Complex cadmium-tellurium oxides open new horizons due to direct bandgap semiconductors' properties and an amorphous phase exhibiting transparent glass-like properties [7, 8]. There exists a pressing challenge for researchers to optimize the synthesis methods of such films.

Conventionally, thin film synthesis methods are categorized into physical and chemical techniques. Physical processes, such as vacuum evaporation and sputtering, require high temperatures and vacuum [9, 10]. In contrast, chemical methods like electrochemical etching [11], deposition [12, 13], the SILAR method (Successive ionic layer adsorption and reaction) [14], sol-gel [15], and others introduce new opportunities for cost-effective, directed synthesis [16, 17]. These methods dictate the surface morphology, crystallinity, and composition of nanostructures, offering avenues for developing new classes of materials [18, 19]. Chemical and electrochemical synthesis methods are pivotal in creating and modifying multi-component and nanostructured materials. Synthesis and tuning component content in compounds are critical, ensuring control over surface and bulk properties [20, 21].

In particular, adjusting the oxidation parameters of CdTe allows obtaining materials with a controllable bandgap width from 1.5 eV (for CdTe) to 3.8 eV, depending on the oxygen concentration [22 - 24]. To date, $CdTe_xO_y$ materials are well described, including CdO (x=0) [25, 26], CdTe (y=0), and various tellurates: CdTeO, Cd_2TeO_4, Cd_3TeO_6, Cd_3TeO_6, $CdTe_3O_8$, $CdTe_2O_5$, $CdTeO_3$ [27-29].

This study is devoted to synthesizing and analyzing $Cd_xTe_yO_z$ films grown via the SILAR method on the CdS/ZnO surface. The synthesis technology was executed through repeating deposition cycles in different ionic electrolyte compositions, where the structural properties of the samples were examined using X-ray diffraction and Raman light scattering techniques.

II. Experiment

A. Substrate Preparation

As a template for the synthesis of $Cd_xTe_yO_z$, CdS layers grown via electrochemical deposition on monocrystalline ZnO wafers were used. A detailed description of the substrate preparation, including the anodic electrochemical reaction and micro-relief formation, is provided in the reference [22]. The choice of CdS layers electrochemically deposited on monocrystalline ZnO wafers as the template for synthesizing $Cd_xTe_yO_z$ was motivated by the high crystalline quality and surface texture of CdS, which is conducive to the controlled growth and phase stability of $Cd_xTe_yO_z$.

B. Synthesis Method

The synthesis method for $Cd_xTe_yO_z$ oxides on a CdS/ZnO substrate involved SILAR. This process comprised cyclical treatment in precursor solutions and removal of excess reagents. The number of treatment cycles was 5 (for the first batch of samples) or 10 (for the second batch of samples), with each process consisting of 4 stages, each having its respective soaking time. The total soaking time in both precursors for both sample batches was consistent and amounted to 100 minutes. Synthesis parameters were chosen to study their effects on the morphology and structural properties of the samples.

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

C. Precursors

The synthesis process of CdS/ZnO heterostructures requires appropriate precursors. Sodium telluride solution with a concentration of 0.01M Na_2TeO_3 served as the tellurium source during the synthesis.

The source of cadmium was an alcoholic solution of cadmium nitrate with a concentration of 0.01M $Cd(NO_3)_2$.

Between the cleaning stages of the samples, hydrogen peroxide was used. The employment of this substance ensures the efficient removal of contaminants and residue from the sample surface.

D. Experimental Conditions

Table 1 describes the experimental conditions for synthesizing $Cd_xTe_yO_z$ oxides on the CdS/ZnO substrate.

TABLE I. EXPERIMENTAL CONDITIONS FOR THE SYNTHESIS OF $Cd_xTe_yO_z$ OXIDES ON THE CdS/ZnO SUBSTRATE

Element	Time, min	
	1st batch of samples	2nd batch of samples
1. Sample immersion in 0.01M Na_2TeO_3 precursor	10	5
2. Rinsing samples in H_2O_2	2	1
3. Sample immersion in 0.01M $Cd(NO_3)_2$ precursor	10	5
4. Rinsing samples in H_2O_2	2	1

Thus, both batches of samples were subjected to precursor treatment for 100 minutes, and rinsing in H_2O_2 took 20 minutes, making the total processing time 120 minutes for each set of samples.

The solution was stirred using a magnetic stirrer at a reduced speed for optimal adsorption and to ensure electrolyte adhesion during the deposition process.

After completing all SILAR cycles, samples underwent thermal treatment in a JetFirst diffusion furnace. The annealing procedure lasted 20 minutes at a temperature of 150°C. The annealing in ambient conditions consolidated surface states through oxygen saturation. Subsequently, samples were naturally allowed to age in the open air for three months.

E. Characterization

The morphological properties of the synthesized structures were investigated using an SEO-SEM Inspect S50-B scanning electron microscope. Surface chemical composition analysis was carried out by energy-dispersive X-ray spectroscopy (EDX) on an AZtecOne device equipped with an X-MaxN20 detector. X-ray diffraction measurements were conducted on a Dron-3M device, using unfiltered Cu Kα radiation in a 2ϑ angle range from 10° to 80° with a step of 0.01°. Raman spectra were acquired using the RENISHAW inVia Reflex system, utilizing an excitation wavelength of 532 nm and an intensity of 5%.

For a detailed analysis of morphological and structural parameters, software tools ImageJ and Vesta were employed, and references were made to crystallographic structures in the Crystallography Open Database (COD).

III. Results

A. SEM Analysis

Figure 1 shows the SEM micrographs of the surface of samples from the first batch (sample 1) and the second batch (sample 2). It can be observed that both samples exhibit the formation of a densely packed globular structure. The surface is not uniform, with occasional "cloudy" regions representing a fluffy, porous structure.

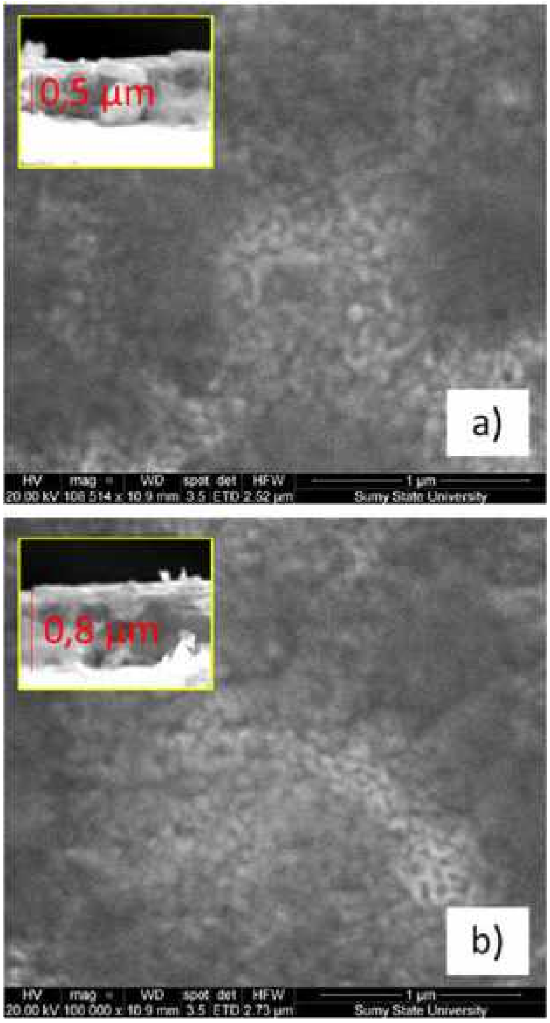


Fig. 1. SEM images of the synthesized $Cd_xTe_yO_z$ surface: a) sample 1; b) sample 2. Insets show the cross-section of these structures.

The insets in Figure 1 display the cross-section of these structures. The layer thickness of the sample from the first batch is 0.5 µm, while for the sample from the second batch, it's 0.8 µm. This variation in thickness can be attributed to the number of cycles (5 versus 10) and the soaking time at each stage. On the other hand, the doubled treatment time at each stage for batch 1 might lead to more stable growth, reflected in a thinner layer (0.5 µm) compared to batch 2 (0.8 µm).

Increasing the number of cycles for batch 2 might result in the gradual accumulation of material, increasing layer thickness. Conversely, a shorter treatment time at each stage might produce a more irregular and dynamic structure. These observations confirm the potential for precise control over surface morphology and layer thickness by adjusting the number of cycles and soaking time at each stage.

B. *EDX Analysis*

EDX spectroscopy was utilized to analyze the chemical compositions of the two samples from different batches (Fig. 2, Table 2).

It can be observed that both samples exhibit spectra from Zn, S, Cd, Te, and O. The minor presence of Zn and S suggest that $Cd_xTe_yO_z$ densely covers the template surface. For both samples, there's a high concentration of oxygen atoms compared to other elements (Table 3), suggesting oxygen in various compounds on the surface.

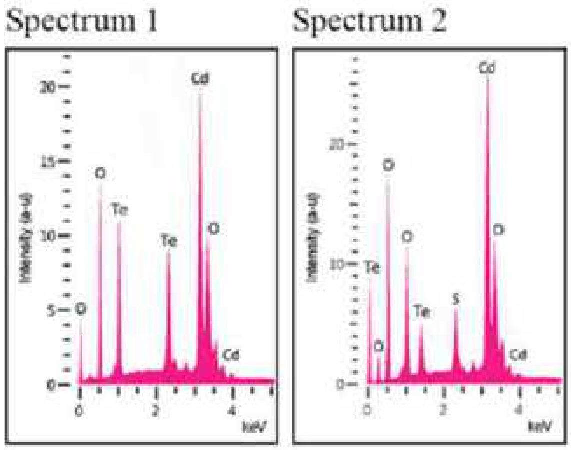


Fig. 2. EDX spectra of synthesized Cd_xTeyO_z.

TABLE II. COMPONENT COMPOSITION ON THE SURFACE OF $Cd_xTe_yO_z$

Samples	At, %				
	Te	O	Zn	S	Cd
1	28.76	38.37	0.13	1.32	31.42
2	25.32	39.41	0.02	2.37	32.88

TABLE III. RATIO OF COMPONENTS ON THE SURFACE $Cd_xTe_yO_z$ HETEROSTRUCTURE $Cd_xTe_yO_z$

Samples	Value			
	Cd/Te	Cd/O	Te/O	(Cd+Te)/O
1	1.09	0.82	0.75	1.57
2	1.30	0.83	0.64	1.47

The Cd/Te and Cd/O ratios for both samples have distinct differences. For sample 1, the Cd/Te ratio is 1.09, whereas, for sample 2, this ratio increases to 1.30. This suggests a higher cadmium concentration in sample 2 relative to tellurium. The Cd/O ratio remains nearly constant between the two samples, indicating a consistent amount of oxygen in the $Cd_xTe_yO_z$ structure. On the other hand, the Te/O ratio is 0.75 and 0.64 for the first and second samples, respectively, pointing to a decrease in tellurium relative to oxygen in sample 2.

These observations can be attributed to variations in the synthesis conditions between the two sample batches. Doubling the number of cycles and reducing the soaking time at each stage for the second batch might influence the chemical composition due to changes in reaction kinetics, deposition mechanisms, and element diffusion. These distinctions might subsequently impact the physicochemical properties and applications of the resultant materials, making these observations crucial for further research and synthesis process optimization. For instance, during shorter treatment cycles, the reaction might not reach equilibrium, leading to incomplete deposition of certain elements. Moreover, short cycles could foster the formation of more dispersed sediment particles, which can alter the layer's thickness

C. *XRD Analysis*

Figure 3 presents the diffractometric spectra of samples 1 and 2, respectively. A perfect alignment of these peaks can be observed. For both samples, the prominent peaks at 2θ = 26.3° for the first sample and 2θ = 26.5° correspond to the (310) plane. There are also intense peaks at 2θ =33.0° and 2θ =33.6° for the first and second samples, respectively, corresponding to the (020) planes. It can be seen that the peaks representing reflections from the (310) plane are characteristic of the cadmium-tellurium oxide family $Cd_xTe_yO_z$, while the peaks showing reviews from the (020) plane are typical of tellurium oxides TeO_z (Fig. 4).

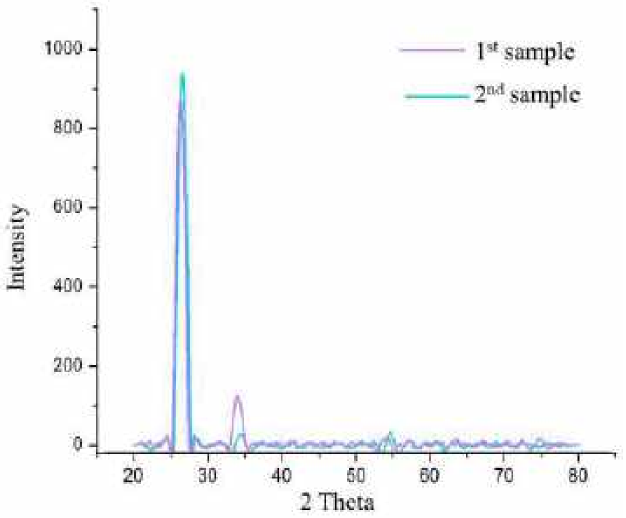


Fig. 3. XRD spectrum $Cd_xTe_yO_z$.

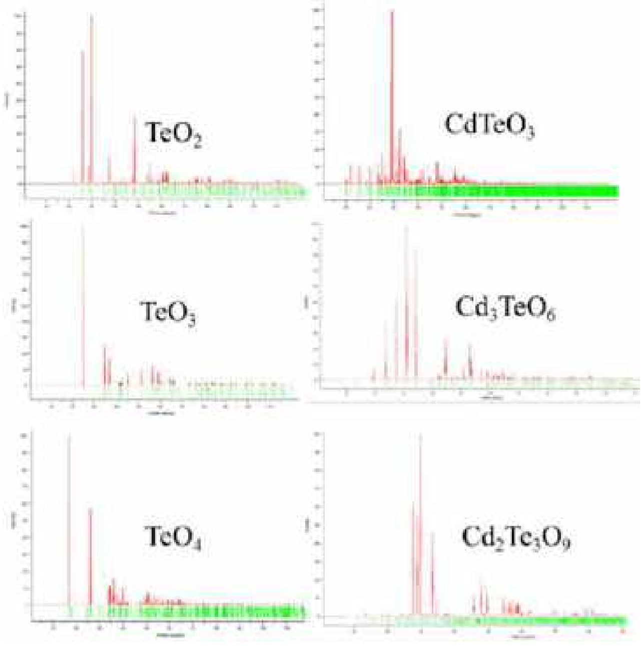


Fig. 4. Overlaid reference spectra TeO_2, TeO_3, TeO_4, $CdTeO_3$, Cd_3TeO_6, $Cd_2Te_3O_9$ from the Crystallography Open Database (COD) visualized using the VESTA program.

The peak corresponding to the reflection from the (310) plane is more intense for sample 2, while the peak from the

(020) plane is more intense for sample 1. Thus, long-cycle treatment can lead to the formation of tellurium oxides, while short cycles result in the formation of $Cd_xTe_yO_z$.

IV. Discussion

The operating principle of SILAR is grounded on the adsorption of ions from a solution onto a surface, initiating a reaction and forming a film. This process largely depends on the chosen precursor and processing conditions. A buffer layer can be employed to enhance the film's adhesion to the substrate. In our study, the CdS layer promoted better formation of cadmium-telluride oxide films.

The family of materials represented by $Cd_xTe_yO_z$ offers a fascinating example of complex oxides where properties can be tailored by altering the parameter 'x'. This parameter denotes the ratio of cadmium to tellurium in the oxide structure and allows for the "fine-tuning" of the material's electronic and optical characteristics.

In the extreme case where x=1, the scenario corresponds to the composite containing only cadmium atoms. The material exhibits typical cadmium oxide (CdO) characteristics in this situation. CdO is a high electrical conductivity conductor generally transparent in the ultraviolet range.

On the opposite end of the spectrum, when x=0, the material consists solely of tellurium atoms. This indicates the properties of tellurium oxide (TeO_2 or another form, depending on the oxidation state). Tellurium oxides are typically dielectrics, possess a high refractive index, and are employed in optics.

Between these two extremes lies a myriad of transitional states of $Cd_xTe_yO_z$, each possessing its unique properties (Table 4). By varying the 'x' parameter, materials can be derived with diverse characteristics, from conductive to dielectric, making this group of oxides a subject of intense investigation.

TABLE IV. CRYSTAL LATTICE PARAMETERS OF THE $Cd_xTe_yO_z$ FAMILY OF MATERIALS

Characteristic	**TeO_2**	**TeO_3**	**TeO_4**
Crystal system	Orthorhombic	Trigonal	Monoclinic
Space group	Pbca	R-3c	P2/c
Space group number	61	167	14
Volume of cell Å^{-3})	395.35	97.26	136.77
a, Å	5.6	5.285	4.96
b, Å	5.75	5.285	5.23
c, Å	12.3	5.285	5.77
α, °	90	57.051	65.83
β, °	90	57.051	90
γ, °	90	57.051	90
Characteristic	**$CdTeO_3$**	**Cd_3TeO_6**	**$CdTe_3O_8$**
Crystal system	Orthorhombic	Trigonal	Monoclinic
Space group	Pnma	R-3h	P2/c
Space group number	62	148	13
Volume of cell Å^{-3})	1196.38	805.01	771.50
a, Å	7.458	9.162	14.066
b, Å	14.522	9.162	5.872
c, Å	11.046	11.0736	10.521
α, °	90	90	90
β, °	90	90	117.4
γ, °	90	120	90

Investigating the crystalline structures of cadmium and tellurium oxides unveils intriguing avenues for further research. Shared features and differences in their crystal systems could provide insights into their unique physicochemical properties and potential applications across various scientific and technological domains.

Conclusions

During the deposition of ionic layers on the CdS/ZnO heterostructure, a dense nanocomposite layer of $Cd_xTe_yO_z$ was successfully formed.

The study revealed that processing conditions influence structural and phase properties. Specifically, increasing the number of cycles can lead to material accumulation and an increase in layer thickness. On the other hand, shorter processing times can result in a more irregular structure. Differences in peak intensities corresponding to various planes further substantiate that processing regimes significantly affect phase and structural features.

Overall, the results affirm that the SILAR method can be pivotal in designing and optimizing nanocomposite materials with precisely defined characteristics for various applications.

One of the key achievements of this study is the identification and quantitative assessment of changes in structural and phase properties, which depend on variations in processing cycles and duration. These findings could pave the way for optimized protocols, further enhancing the quality of films fabricated through the SILAR method.

Addressing the concerns regarding the reproducibility of data, we assert that the distinct diffraction peak intensities observed in XRD analysis signify a controlled modulation of the film's structure. Despite the noted inhomogeneities, the consistent trends across different batches underscore reproducibility in the data, demonstrating the potential of the SILAR method in achieving specific desired characteristics through careful manipulation of processing parameters.

Furthermore, other prevalent methods may yield superior homogeneity and optical quality films. However, the SILAR method's cost-effectiveness and more straightforward equipment requirements stand out. Additionally, this method facilitates layer-by-layer assembly, which grants a remarkable degree of control over the film thickness and composition at the nanoscale level. Notably, the perceived inhomogeneities and porosity might engender enhanced surface area and reactivity, making these films particularly potent in applications such as photocatalysis and sensor development, where such characteristics are sought-after.

To further consolidate the advantages of the SILAR method, future studies could venture into optimizing the processing conditions to mitigate the observed inhomogeneities, thereby establishing a balance between cost-effectiveness and material quality.

Acknowledgment

The study was supported by the Ministry of Education and Science of Ukraine via Project No. 0122U000129 “The search for optimal conditions for nanostructure synthesis on the surface of A3B5, A2B6 semiconductors and silicon for photonics and solar energy” and Project No. 0121U10942 “Theoretical and methodological bases of system fundamentalization of the future nanomaterials experts training for productive professional activity”. In addition, the research was partly supported by COST Action CA20129 “Multiscale Irradiation and Chemistry Driven Processes and

Related Technologies" (MultIChem), and COST Action CA20126 - Network for research, innovation, and product development on porous semiconductors and oxides (NETPORE). A.P. thanks to the Institute of Solid-State Physics, University of Latvia. ISSP UL as the Center of Excellence is supported through the Framework Program for European universities, Union Horizon 2020, H2020-WIDESPREAD-01–2016–2017-TeamingPhase2, under Grant Agreement No. 739508, CAMART2 project..

References

[1] H. Klym, I. Karbovnyk, A. Luchechko, Y. Kostiv, V. Pankratova, and A.I. Popov, "Evolution of free volumes in polycrystalline BaGa2O4 ceramics doped with Eu3+ ions," Crystals, vol. 11, no. 12, pp. 1515, 2021.

[2] H. Klym, A. Ingram, O. Shpotyuk, O. Hotra, and A.I. Popov, "Positron trapping defects in free-volume investigation of Ge–Ga–S–CsCl glasses," Radiation Measurements, vol. 90, pp. 117-121, 2016.

[3] V.P. Savchyn, A.I. Popov, O.I. Aksimentyeva, H. Klym, Y.Y. Horbenko, V. Serga, A. Moskina, and I. Karbovnyk, "Cathodoluminescence characterization of polystyrene-BaZrO3 hybrid composites," Low Temperature Physics, vol. 42, no. 7, pp. 760-763, 2016.

[4] H. Klym, I. Karbovnyk, M.C. Guidi, O. Hotra, and A.I. Popov, "Optical and Vibrational Spectra of CsCl-Enriched GeS2-Ga2S3 Glasses," Nanoscale Research Letters, vol. 11, no. 1, art. no. 132, 2016.

[5] S. Yana, "Porous indium phosphide: Preparation and properties," in Handbook of Nanoelectrochemistry: Electrochemical Synthesis Methods, Properties, and Characterization Techniques, 2016, pp. 283–306.

[6] A. Usseinov, Z. Koishybayeva, A. Platonenko ... and A.I. Popov, "Ab-Initio Calculations of Oxygen Vacancy in Ga2O3 Crystals," Latvian Journal of Physics and Technical Sciences, vol. 58, no. 2, pp. 3–10, 2021.

[7] H. Arizpe - Chávez, R. Ramírez - Bon, F. J. Espinoza - Beltrán, O. Zelaya - Angel, J. González - Hernández, and L. Baños, "Optical and structural properties of CdTe - $CdTeO_3$ nanocrystalline composites," in AIP Conference Proceedings, vol. 378, no. 1, pp. 203-209, July 1996.

[8] Y. Suchikova, S. Kovachov, I. Bohdanov, ..., A. Moskina, and A. Popov, "Characterization of $Cd_xTe_yO_z$/CdS/ZnO Heterostructures Synthesized by the SILAR Method," Coatings, vol. 13, no. 3, pp. 639, 2023.

[9] P. Aabel, A. Anupama, and M. S. Kumar, "Preparation and characterization of CZTS thin films by vacuum-assisted spray pyrolysis and fabrication of Cd-free heterojunction solar cells," Semiconductor Science and Technology, vol. 38, no. 4, p. 045010, 2023.

[10] P. Abraham, S. Shaji, D. A. Avellaneda, J. A. Aguilar-Martínez, and B. Krishnan, "(002) oriented ZnO and ZnO: S thin films by direct ultrasonic spray pyrolysis: A comparative analysis of structure, morphology and physical properties," Materials Today Communications, vol. 35, p. 105909, 2023.

[11] Y.A. Suchikova, V.V. Kidalov, and G.A. Sukach, "Influence of the carrier concentration of indium phosphide on the porous layer formation," Journal of Nano- and Electronic Physics, vol. 2, no. 4, pp. 75–81, 2010.

[12] S. K. Hwang, I. J. Park, S. W. Seo, J. H. Park, S. J. Park, and J. Y. Kim, "Electrochemically deposited CZTSSe thin films for monolithic perovskite tandem solar cells with efficiencies over 17%," Energy & Environmental Materials, e12489, 2023.

[13] Y. Suchikova, S. Kovachov, and I. Bohdanov, "Formation of oxide crystallites on the porous GaAs surface by electrochemical deposition," Nanomaterials and Nanotechnology, vol. 12, 2022.

[14] M. Sathya, G. Selvan, M. Karunakaran, K. Kasirajan, S. Usha, M. Logitha, ..., and P. Baskaran, "Synthesis and characterization of cadmium doped on ZnO thin films prepared by SILAR method for photocatalytic degradation properties of MB under UV irradiation," The European Physical Journal Plus, vol. 138, no. 1, pp. 1-12, 2023.

[15] K. Chehhat, A. Mecif, A. H. Mahdjoub, R. Nazir, M. A. Pandit, F. Salhi, and A. Noua, "Sol-gel synthesis of porous cobalt-doped ZnO thin films leading to rapid and large scale Orange-II photocatalysis," Journal of Sol-Gel Science and Technology, vol. 106, no. 1, pp. 85-94, 2023.

[16] Y. Suohikova, S. Vambol, V. Vambol, N. Mozaffari, and N. Mozaffari, "Justification of the most rational method for the nanostructures synthesis on the semiconductors surface," Journal of Achievements in Materials and Manufacturing Engineering, vol. 92, no. 1-2, 2019.

[17] L. Xiang, Q. Li, C. Li, Q. Yang, F. Xu, and Y. Mai, "Block copolymer self - assembly directed synthesis of porous materials with ordered bicontinuous structures and their potential applications," Advanced Materials, vol. 35, no. 5, p. 2207684, 2023.

[18] J. Bashir, M. B. Chowdhury, R. R. Kathak, S. Dey, A. T. Tasnim, M. A. Amin, ..., and M. S. A. Hossain, "Electrochemical fabrication of mesoporous metal-alloy films," Materials Advances, vol. 4, no. 2, pp. 408-431, 2023.

[19] Y. Suchikova, "Provision of environmental safety through the use of porous semiconductors for solar energy sector," Eastern-European Journal of Enterprise Technologies, vol. 6, no. 5, pp. 26–33, 2016.

[20] Kapadnis, R. S., Kale, S. S., & Wagh, V. G. (2013). Studies on chemically synthesis of polycrystalline $CdTeO_3$ thin films. Studies, 3(8).

[21] Lai, Y., Wang, Y., Zhu, Y., Guo, R., Xia, Y., Huang, W., & Li, Z. (2018). Irregular micro-island arrays of CdO/CdS composites derived from electrodeposited Cd for high photoelectrochemical performances. Journal of The Electrochemical Society, 165(3), H91 https://doi.org/10.1149/2.0321803jes

[22] Y. Suchikova, S. Kovachov, I. Bohdanov, ..., V. Pankratov, and A.I. Popov, "Study of the structural and morphological characteristics of the $Cd_xTe_yO_z$ nanocomposite obtained on the surface of the CdS/ZnO heterostructure by the SILAR method," Applied Physics A: Materials Science and Processing, vol. 129, no. 7, pp. 499, 2023.

[23] J. Carmona-Rodríguez, R. Lozada-Morales, O. Jiménez-Sandoval, F. Rodríguez-Melgarejo, M. Meléndez-Lira, and S. J. Jiménez-Sandoval, "$CdTeO_x$ to $CdTeO_3$ structural phase transition in as-grown polycrystalline films by reactive sputtering," Journal of Applied Physics, vol. 103, no. 12, p. 123516, 2008.

[24] F. Caballero-Briones, A. Zapata-Navarro, A. Martel, A. Iribarren, J. L. Peña, R. Castro-Rodríguez, and S. Jiménez-Sandoval, "Compositional mixture in RF sputtered CdTe oxide films. Raman spectroscopy results," Superficies y Vacio, vol. 16, no. 3, pp. 38-42, 2003.

[25] A. Jayaraman and G. A. Kourouklis, "A high pressure Raman study of TeO_2 to 30 GPa and pressure-induced phase changes," Pramana, vol. 36, no. 2, pp. 133-141, 1991.

[26] M. Ceriotti, F. Pietrucci, and M. Bernasconi, "Ab initio study of the vibrational properties of crystalline TeO_2: The α, β, and γ phases," Physical Review B, vol. 73, no. 10, p. 104304, 2006.

[27] A. Chagraoui, I. Yakine, A. Tairi, A. Moussaoui, M. Talbi, and M. Naji, "Glasses formation, characterization, and crystal-structure determination in the Bi_2O_3–Sb_2O_3–TeO_2 system prepared in an air," Journal of materials science, vol. 46, no. 16, pp. 5439-5446, 2011.

[28] İ. Kabalcı, G. Özen, and M. L. Öveçoğlu, "Microstructure and crystallization properties of TeO_2-PbF_2 glasses," Journal of Raman Spectroscopy, vol. 40, no. 3, pp. 272-276, 2009.

[29] A. Guillén-Cervantes, M. Becerril-Silva, H. E. Silva-López, J. S. Arias-Cerón, E. Campos-González, M. Pérez-González, and O. Zelaya-Ángel, "Structural and optical properties of CdTe+$CdTeO_3$ nanocomposite films with broad blueish photoluminescence," Journal of Materials Science: Materials in Electronics, vol. 31, no. 9, pp. 7133-7140, 2020.

Modeling of Ion Transfer Processes in $CsPbBr_3$ Crystals

Ihor Kayun
Department of Sensor and Semiconductor Electronics
Ivan Franko National University of Lviv
Lviv, Ukraine
ihor.kaiun@lnu.edu.ua

Roman Lys
Department of Sensor and Semiconductor Electronics
Ivan Franko National University of Lviv
Lviv, Ukraine
roman.lys@lnu.edu.ua

Yuriy Tymkiv
Department of Sensor and Semiconductor Electronics
Ivan Franko National University of Lviv
Lviv, Ukraine
y.tymkiv.official@gmail.com

***Abstract* — In the research paper, an analysis of voids and channels in the crystal structure of $CsPbBr_3$ is carried out with the help of ToposPro 5.5.2.0 – the automated stereo atomic crystal structure analysis system for determining significant voids and channels. The Voronoi-Dirichlet partition is used to tackle these crystal-chemical tasks. The geometric characteristics of the voids and channels are calculated. It is found that all elementary voids are significant for $CsPbBr_3$, but the channels that can pass Cs^+ and Pb^{2+} ions do not meet the requirements. The conductivity chains of bromine ions are observed at room temperature, which may indicate ionic conductivity if the temperature increases due to the migration of anions.**

Keywords — perovskites, $CsPbBr_3$, ionic conductivity, Voronoi-Dirichlet partition.

I. Introduction

In recent years, inorganic halides of perovskites have drawn increased attention among the scientific community [1-3]. These materials have beneficial properties for use in emitting diodes as laser-active media, photosensitive elements for solar cells, etc. [4, 5]. However, before the widespread introduction of $CsPbBr_3$ into practical use, it is necessary to study its electro-physical properties in detail, such as conductivity, resistance, and possible charge carrier transfer mechanisms. It will facilitate the development of effective synthesis methods and optimize crystal growth conditions, ensuring high quality and stability of electrical and optical properties.

$CsPbBr_3$ has a perovskite crystal structure characterized by Cs^+ and Pb^{2+} ions located in the center of Br_6 octahedra. This structure determines the fundamental physical properties of the crystal, such as conductivity and photosensitivity.

$CsPbBr_3$ has high electrical conductivity at room temperature, which ensures its use in various electronic devices. Electrical conductivity depends on the concentration of defects in the crystal and on the temperature. As the temperature rise, electrical conductivity increases. Popular studies, which consider high ionic conductivity in single crystals of these materials, explain this by the migration of anions [6]. Another study also proves this, suggesting that these materials are halide-ion conductors and that their ionic conductivities are close to other well-known halide-ion $PbBr_2$ and $PbCl_2$ conductors [7]. The migration activation energy calculated in this study was 0.25 *eV* for $CsPbBr_3$. In particular, this study clarifies that the conductivity is due to anion migrations.

Compounds with a perovskite structure easily change their crystal structure from orthorhombic to cubic through tetragonal. In addition, the coexistence of several phases is often observed in certain temperature ranges. A phase transformation from orthorhombic to cubic occurs in the temperature range between room temperature and 473 *K*. This observation is confirmed by other studies on this topic, in which other phase transformations are also described [8], namely a tetragonal phase is distinguished between orthorhombic and cubic.

The research attempted to study phase transformations in $CsPbBr_3$ single crystal using a direct current [6]. Measurements were made between 30 and 380 °*C*. A noticeable increase in conductivity is observed at 90 °*C*, which corresponds to the orthorhombic phase transformation from tetragonal. There are also other studies of phase transitions in this crystal. Cola et al. found a phase transition only at 123 °*C* with an insignificant thermal effect (less than 0.7 $J \cdot g^{-1}$) [9]. Through a detailed study, phase transitions of the second and first orders at 88 °*C* and 130 °*C*, respectively, were determined in the works of Hirotsu et al. [10, 11].

A thorough study of the possibility of using $CsPbBr_3$ crystals as solid electrolytes was conducted [7, 12]. For this purpose, the authors measured the ion conductivity and ion transport number of $CsPbBr_3$. Conductivity measurements were limited by the melting of the samples in some experiments. Thus, found melting points are 500 ± 10 °*C* for $CsPbBr_3$. For comparison, the melting point of $PbBr_2$ is 371 °*C* [13]. Due to the high melting point of $CsPbBr_3$, the conductivity of $CsPbBr_3$ was higher than that of $PbBr_2$.

II. Method

A $CsPbBr_3$ single crystal grown by the Stockbarger method was used for research. The synthesis was carried out in a quartz ampoule with high purity CsBr and $PbBr_2$. Then the obtained single crystal was chipped into thinner samples for further experimental studies. For electrical measurements, ohmic contacts based on silver paste were applied.

Overall, the conductive properties of $CsPbBr_3$ crystals have not been sufficiently studied. However, in these ionic compounds, the conductivity can be theoretically calculated by calculating the Voronoi-Dirichlet polyhedra and constructing crystal conductivity maps at different temperatures. The Voronoi-Dirichlet partition is used to tackle these crystal-chemical tasks because it helps to obtain a map of the system of voids and channels for structures in the form of three-dimensional graphs: an atomic mesh and a

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

cavity mesh, and the edges of the meshes correspond to interatomic distances or probable channels in the structure of the compound.

Voids and channels in crystal structures were analyzed by the ToposPro 5.5.2.0 automated stereo atomic crystal structure analysis system in this work. The search for voids and channels using the Dirichlet program included in the Topos software package consisted of the following stages:

1. Construction of the Voronoi-Dirichlet partition (VDP) of the crystal space for all independent atoms of the structure, which includes the formation of the Voronoi-Dirichlet partition.

2. Determination of the vertices coordinates of atomic VDPs and elementary voids.

3. Determination of the atoms' VDP edges and all elementary channels.

4. Calculation of geometric characteristics of elementary voids and channels.

The results of the calculations are stored in the form of a three-level adjacency matrix of the Voronoi-Dirichlet partition, where the first level contains information about voids and the second level – about channels that connect voids with neighboring ones. The obtained data make it possible to determine significant voids and channels, as well as analyze ion conductivity. Ionic conductivity in $CsPbBr_3$ crystals is calculated for ions based on the obtained geometric characteristics of voids and channels. Literature data on crystal structure, ion radii, and other parameters necessary for the analysis of ionic conductivity were used in the calculations.

III. The Results

We have experimentally confirmed that the conductivity of $CsPbBr_3$ crystals is practically the same along different crystallographic directions.

Also, it was experimentally established that the conductivity of $CsPbBr_3$ single crystals increases with increasing temperature. By rearranging the spectra of thermally stimulated conductivity in logarithmic coordinates (Fig. 1), the activation energy of the conductivity process was determined to be 0.3 *eV*. Fig. 1 shows the curves of thermally stimulated conductivity, on which the current was measured along the direction that is perpendicular to the plane of easy chipping of the crystals.

We conducted an experiment on the $CsPbBr_3$ crystal and found that the long-term effect of constant voltage on the crystal led to its coloring and increased resistance. The crystal had a visible darkening from the cathode side, which spread inside the crystal. The darkening extended to the entire length of the crystal after the experiment was stopped.

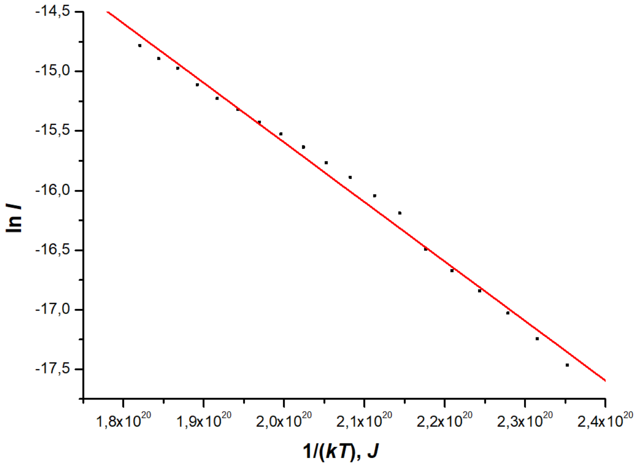


Fig. 1. Calculation of the conductivity activation energy of $CsPbBr_3$ crystals

The appearance of alterations in the crystal may be associated with a change in the material's structure under the influence of an electric field. It is known that an electric field can affect the position and movement of ions in a crystal, which can lead to changes in the electrical and optical properties of the material. The increase in resistance of the crystal, which was found in the experiment, may be associated with a change in the concentration of charge carriers in the crystal.

Literature structural data were used, namely lattice parameters and atomic positions to construct the conductivity map of $CsPbBr_3$ [14]. Significant voids were calculated at 4 *K*, room temperature, 473 *K*, and 773 *K* for Br^-, Pb^{2+}, and Cs^+ ions. The radii of Br^- (1.82 *Å*), Pb^{2+} (1.33 *Å*), and Cs^+ (1.81 *Å*) ions were used for the calculation and analysis of elementary voids.

According [14] below 473 *K* crystal structure of $CsPbBr_3$ is orthorhombic with orthorhombic distorted perovskite structure (Fig. 2).

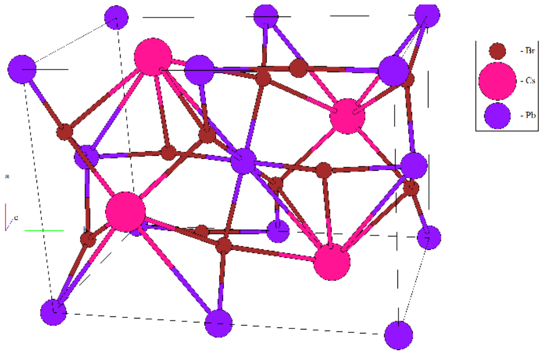


Fig. 2 Crystal structure of $CsPbBr_3$ at 298 *K*. Lattice parameters is: *a* = 0.82446 *nm*, *b* = 1.17399 *nm*, *c* = 0.81915 *nm* [14]

At 473-773 *K* crystal structure has cubic perovskite structure (Fig. 3).

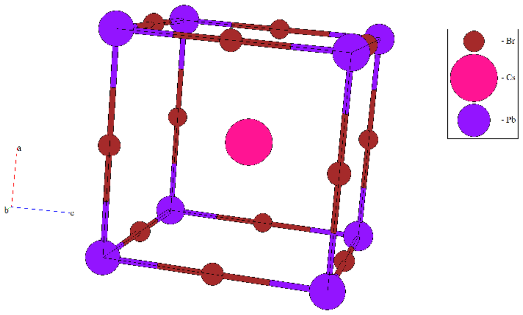


Fig. 3 Crystal structure of $CsPbBr_3$ at 773 *K*. Lattice parameter is: $a = 0.59281$ *nm* [14]

We will define some parameters that were used to model the ionic conductivity and construct the conductivity map of the $CsPbBr_3$ crystal.

An elementary void is a region of the crystal space, the center of which is one of the VDP peaks of one of the atoms. Atoms forming an elementary cavity are called atoms whose VDP converges in the center of this elementary void. The center of the void can be located both inside and outside of the polyhedron formed by the atoms forming the cavity. Basic and non-basic elementary voids are distinguished, which we will denote by *ZA* and *ZC*, and which form the set {*ZA*} and {*ZC*} with the assigned ordinal numbers *N* of voids of the corresponding type *ZAN* and *ZCN*.

The radius of the elementary void is the radius of the sphere whose volume is equal to the volume of the VDP elementary cavity. Physically, the radius of an elementary cavity corresponds to the radius of an atom that can fit into the void, considering the influence of the crystal field.

An elementary channel is a channel connecting two elementary voids. It corresponds to the VDP edge of any of the atoms forming both voids. Such an edge is called an elementary channel line. Atoms forming an elementary channel are called atoms whose VDP has a common edge that coincides with the line of the elementary channel.

An atom can pass through an elementary channel if the sum of its radius (r_i) and the average radius of the atoms forming the channel (r_a) do not exceed the cross-sectional radius of the channel (r_c). A deformation coefficient $\gamma \leq 1$ is introduced to consider the possible polarization (deformation) of ions when they pass through the channel. Then the specified condition is written: $\gamma \cdot (r_i + r_a) \leq r_c$. The value of γ depends on the nature of mobile cations and anions of the framework.

A significant elementary void and a significant elementary channel – a void and a channel available for particles considered within the framework of a specific crystal chemical problem. They are the ones that have an obvious physical meaning.

A significant elemental void and a significant elemental channel are considered probable if migration through this path is difficult for one reason or another. The criteria for determining probabilistic elementary voids and channels also depend on the specifics of the certain problem.

The calculation of the adjacency matrix carried out for Cs atoms at temperatures of 4 *K*, room temperature, 473 *K*, and 773 *K* showed that all elementary voids are significant. Considering the deformation coefficient of 0.95, it was found that channels with a radius ≥ 3.44 *Å* would be significant. The obtained channels do not meet this condition (Figs. 4, 5).

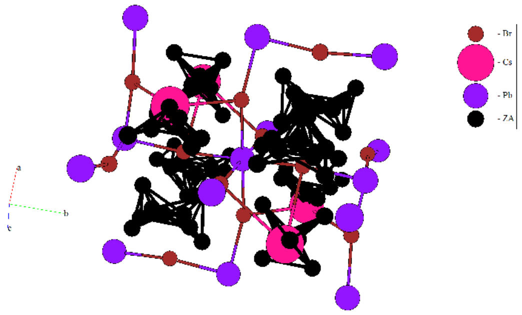


Fig. 4. Calculated channels for Cs^+ ions at room temperature (including not significant)

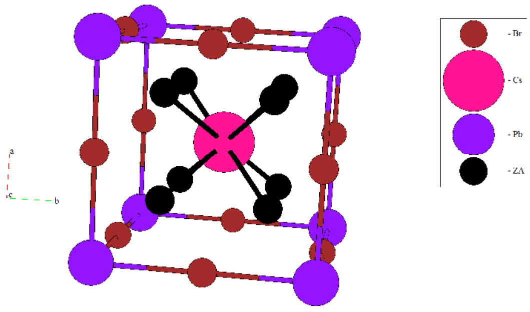


Fig. 5. Calculated channels for Cs^+ ions at a temperature of 773 *K* (including not significant)

The calculation performed for Pb^{2+} atoms showed that all elementary voids are significant, and the channels that can pass the Pb^{2+} ion should be the channels with a radius greater than $0.95 \cdot (1.33 + 1.82) = 2.99$ *Å*. The obtained channels also do not meet this condition (Figs. 6, 7).

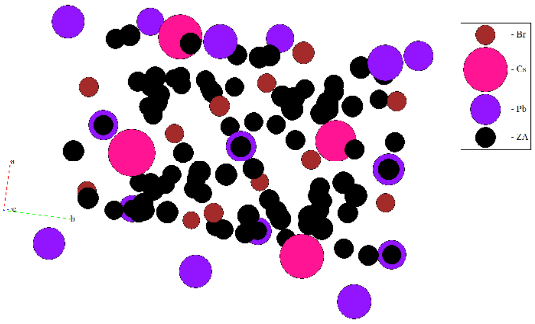


Fig. 6. Calculated set of voids for Pb^{2+} ions at room temperature (some of them overlap with Pb positions)

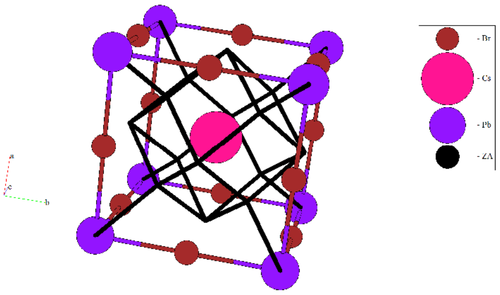


Fig. 7. Calculated channel for Pb^{2+} ions at a temperature of 773 *K* (that are not significant)

The calculation for Br-atoms showed that all eight elementary voids are significant, and channels with a radius greater than 2.99 Å will be the channels that can pass the Br$^-$ ion. At 4 *K*, all channels will be significant except the voids *ZA*1, which are associated with *ZA*2 and *ZA*6 because they are probabilistic channels (Fig. 8).

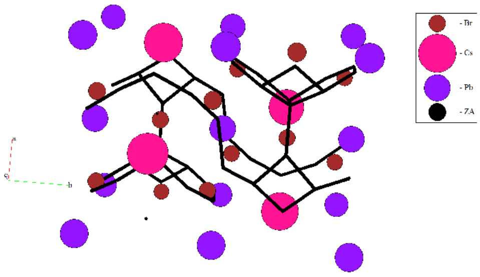


Fig. 8. Significant channels are calculated for Br$^-$ ions at a temperature of 4*K*

The channels associated with all voids are significant at room temperature and subsequent temperature increases (Fig. 9, 10, 11). Thus, ionic conductivity in $CsPbBr_3$ crystals is unlikely at a temperature of 4 *K*. The conductivity chains of bromine ions are already observed at room temperature, which can form ionic conductivity with further temperature increase.

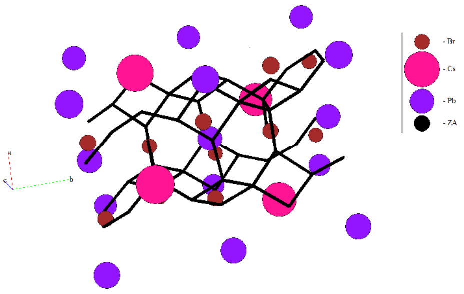


Fig. 9. Significant channels are calculated for Br$^-$ ions at room temperature

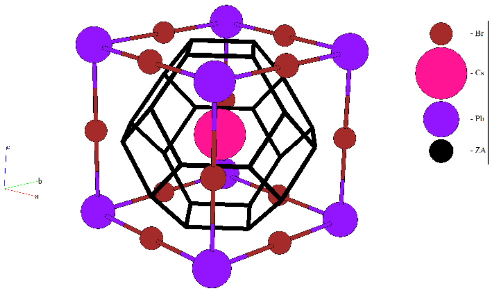


Fig. 10. Conductivity map is calculated for Br$^-$ ions at a temperature of 473 *K*

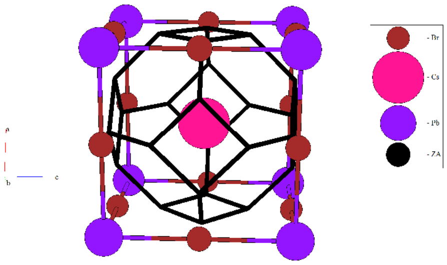


Fig. 11. Significant channels are calculated for Br$^-$ ions at a temperature of 773 *K*

Conclusions

Existing studies have shown that $CsPbBr_3$ has several phase transitions at high temperatures. It was found that the electrical conductivity of these crystals is ionic, which can be theoretically calculated. Conductivity measurements show dependence on the direction and electric field. The calculated activation energy of the conduction process is 0.3 *eV*.

The results indicate that ionic conductivity in $CsPbBr_3$ crystals is unlikely at low temperatures. However, conductivity chains can appear as the temperature increases, which contributes to anionic conductivity in the studied crystals.

The obtained results are an important step in understanding the electrical conductivity of $CsPbBr_3$, but additional studies are still required to reveal the potential of $CsPbBr_3$ and use it in practice.

References

[1] L.-I. I. Bulyk, O. T. Antonyak, Ya. M. Chornodolskyy et al, “Conductivity of $CsPbBr_3$ at ambient conditions” Journal of Physical Studies, vol. 4, 2021, pp. 4801-1 - 4801-5.

[2] C. Chen, Q. Fu, P. Guo et al, “Ionic transport characteristics of large-size $CsPbBr_3$ single crystals” Res. Express, №6, 2021.

[3] P. Pal, A. Ghosh, “Ionic conduction and relaxation mechanisms in three-dimensional $CsPbCl_3$ perovskite” J. Appl. Phys., №129, 2021.

[4] Y. Fujii, S. Hishino, Y. Yamada and G. Shirane, “Neutron-scattering study on phase transitions of $CsPbCl_3$” Phys. Rev. B9., 1974, p. 4549.

[5] H. Ohta, J. Harada, S. Hirotsu, “Superstructure and phase transitions in $CsPbCl_3$” Solid State Commun., vol. 13, 1973, pp. 1969-1972.

[6] R. Lakshmi Narayan, M. V. S. Sarma, S. V. Suryanarayana, “Ionic conductivity of $CsPbCl_3$ and $CsPbBr_3$” Journal of materials Science Letters, №6, 1987, pp. 93-94.

[7] Ju. Mizusaki, Kimiyasu Arai and Kazuo Fueki, “Ionic conduction of the perovskite-type halides” Solid State Ionics, №11, 1983, pp. 203-211.

[8] C. C. Stoumpos, “Crystal Growth of the Perovskite Semiconductor $CsPbBr_3$: A New Material for High-Energy Radiation Detection” Department of Chemistry, №13, 2013, pp. 2722-2727.

[9] M. Cola, V. Massarotti, R. Riccardi et al, “Binary Systems Formed by Lead Bromide with (Li, Na, K, Rb, Cs and Tl) Br: a DTA and Diffractometric Study” Zeit. Naturforsch, №26, 1971, pp. 1328-1332.

[10] S. Hirotsu, J. Harada, M. Iizumi et al, “Structural Phase Transitions in CsPbBr3” Journal of the Physical Society of Japan, №37, 1974, pp. 1393-1398.

[11] S. Hirotsu, T. Suzuki and S. Sawada, “Ultrasonic Velocity around the Successive Phase Transition Points of $CsPbBr_3$” Journal of the Physical Society of Japan, №43, 1977, pp. 575-582.

[12] H. Hoshino, S. Yokose, M. Shimoji, “Ionic conductivity of lead bromide crystals” J. Solid State Chem., №7, 1973, pp. 1-6.

[13] J.A. Dean, Lange's handbook of chemistry, 12th Ed., New York: McGraw-Hill, 1979, 1291 p.

[14] C.A Lopez, Crystal Structure Features of $CsPbBr_3$ Perovskite Prepared by Mechanochemical Synthesis, № 5, 2020, pp. 5931-5938.

Enhanced Nanostructure Evolution in Functional Spinel Ceramics through Additional Phases

Halyna Klym
Department of Specialized Computer Systems
Lviv Polytechnic National University
Lviv, Ukraine
halyna.i.klym@lpnu.ua

Yuriy Kostiv
Department of Information Technology of Security
Lviv Polytechnic National University
Lviv, Ukraine
yura.kostiv@gmail.com

Ivan Hadzaman
Ivan Franko Drohobych State Pedagogical University,
Drohobych, Ukraine
hadzaman@i.ua

Ivan Karbovnyk
Department of Electronics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
ivan.karbovnyk@lnu.edu.ua

Oleksii Kushnir
Department of Electronics and Computer Technologies
Ivan Franko National University of Lviv
Lviv, Ukraine
oleksiy.kushnir@lnu.edu.ua

***Abstract* — The intricate process of converting defects and pores present in spinel-type ceramics, triggered by the incorporation of supplementary phases, unfolds along both the avenues of two-component decomposition pathway. Investigation has validated that a heightened concentration of these supplementary phases within the ceramic matrix precipitates the gradual division of voids. These fragmented voids then progressively engage in agglomeration phenomena, driven by temporal forces. The phases that are extracted in close proximity to the intergranular boundaries play a pivotal role in shaping the material's characteristics. These extracted phases give rise to emergent sites possessing the unique ability to ensnare and confine positrons, forming localized regions within the ceramics that are particularly effective at capturing these elusive particles.**

Keywords* — *ceramics, addition phase, free volume, defects, positron trapping

I. Introduction

Materials with functional applications (e.g. crystals, nanostructures, nanofilms, etc. [1-3]) are promising in many areas and directions, especially ceramic materials with a spinel structure. Spinel-type ceramics have attracted significant attention within the scientific community due to their immense potential for diverse applications in various fields [4-8]. Application of spinel temperature-sensitive $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$ ceramics encompass negative temperature coefficient thermistors, precise temperature sensors, and in-rush current, etc. [9,10]. The versatility of these ceramics has made them a subject of extensive investigation, with researchers striving to uncover their hidden potential and maximize their utility.

One of the critical factors influencing the functionality and reliability of these ceramics is the temperature-time sintering process [11]. Previous research has shed light on its pivotal role in determining the presence and distribution of additional phases within both the bulk and surface of ceramic samples [11,12]. Surprisingly, an intriguing discovery emerged when researchers reduced the content of the NiO phase in these ceramics – it resulted in a remarkable reduction in thermal aging, with observed relative resistance changes not exceeding of 3%. This unexpected outcome underscores the intricacies at play in the behavior of these ceramics and the potential for fine-tuning their properties to achieve desirable outcomes.

In the quest to mitigate degradation effects in ceramics, a common practice is to incorporate chemical modifications through metallic additives during the initial stages of ceramics preparation [11,12]. These strategically placed metallic additives, positioned within intergranular regions near boundaries, play a pivotal role in suppressing thermally-activated aging phenomena. Their presence stabilizes the cationic distribution within individual ceramic grains, resulting in enhanced stability when compared to non-modified ceramics. This approach highlights the intricate chemistry and material science involved in enhancing the performance and longevity of these ceramics.

Nevertheless, the complexity of the structure of these spinel-type ceramics, spanning individual grains, intergranular boundaries, and pores, presents significant challenges to researchers [13]. Progress in this field hinges on the development of novel characterization techniques that can complement traditional methods. This necessity extends to positron annihilation lifetime (PAL) spectroscopy, a relatively recent addition to the arsenal of techniques applicable to fine-grained functional materials [14,15].

PAL spectroscopy, with its high sensitivity to low electron density, offers valuable insights into the distribution of void species within the structural network of solids [9]. However, interpreting PAL data in the context of ceramics is notably challenging. These data are predominantly influenced by the crystallographic attributes of individual grains, while structural anomalies arising from intergranular contacts within ceramics add layers of complexity to the analysis.

The primary objective of this study is to delve deep into the structural intricacies of high-reliability transition-metal manganite $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$ grain-pore interactions. To achieve this, PAL spectroscopy will be employed alongside conventional structural characterization methods. This comprehensive approach seeks to unveil the inner workings of these ceramics, unravel their mysteries, and advance our comprehension of their performance and stability. By doing so, researchers aim to unlock new possibilities for the application of these ceramics in a wide array of technological advancements.

II. Experimental

Precise quantities of high-purity carbonate salts, which had undergone rigorous testing, were meticulously weighed

979-8-3503-8309-6/23/$31.00 ©2023 IEEE

and subjected to wet mixing. Subsequently, this composite underwent thermal decomposition under ambient air conditions at 700 ºC for a duration of 4 hours [16]. The resulting powders were then subjected to milling, blended with an organic binder, and meticulously pressed to form disks, each approximately 10 mm in diameter and 1 mm in thickness. Four distinct batches of these prepared blanks were sintered, with each batch adhering to specific time-temperature regimes described in [16], which depicts the sintering profiles for $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$ ceramics.

It is crucial to highlight that the sintering process of these ceramics was meticulously orchestrated to create the ideal conditions for inhibiting degradation [14]. Of particular significance is the presence of an additional NiO phase with a NaCl structure, which plays a pivotal role in shaping the final ceramic structure. In essence, these ceramics can be characterized as Ni-deficient when compared to the stoichiometric composition of $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$ that serves as the baseline in the disproportionality calculations. Four distinct batches of $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$ ceramics were prepared, each containing varying proportions of NiO phase ranging from 1 to 12%. The differentiation in NiO content was determined based on the amount of thermal energy transferred during the sintering process (sample No 1 – 1% NiO, sample No 2 – 8% NiO, sample No 3 – 10% NiO, sample No 4 – 12% NiO). The latter was numerically calculated as the square value bounded by the temperature-time curve, positioned above the straight line corresponding to 920 °C, which is the temperature at which monophase $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$ ceramics form [16].

Upon microstructure characterization through X-ray diffractometry, it was observed that the lattice constant of the primary spinel slightly increased from 8.38 Å to 8.41 Å. However, there were no significant alterations in the parameters of the additional NiO phase, which remained within the range near 4.18 Å. This held true even as the content of the NiO phase varied from 1 to 8, 10, and 12%.

To delve deeper into the microstructure of the sintered ceramics, electron microscopy employing a JSM-6700F instrument was employed. Cross-sectional morphology analyses were conducted, focusing on samples obtained from near the surface and chip centers.

Subsequent PAL measurements were carried out using an ORTEC spectrometer, with a ^{22}Na source placed between two sandwiched ceramic samples [16,17]. The data obtained were meticulously analyzed using the LT computer program [17], employing a two-component fitting procedure to yield the most accurate results. The computed values for trapping parameters, including positron lifetime in defect-free bulk (τ_b), average positron lifetime ($\tau_{av.}$), and the positron trapping rate of defects (κ_d) were derived from short and long positron-trapping lifetimes (τ_1 and τ_2), along with component intensities (I_1 and $_{I2}$, where $I_1 + I_2 = 1$) [14]. Moreover, the difference (τ_2 - τ_b) was employed as a metric for assessing the size of extended defects where positrons are ensnared, while the τ_2/τ_b ratio provided insights into the nature of these defects [16,17].

III. Results and Discussion

To elucidate the aforementioned phenomena, an in-depth analysis of the microstructure of the prepared ceramics was conducted. As illustrated in Fig. 1, the ceramics samples prepared exhibit pronounced variations in their grain-pore microstructure.

Sample No 1 showcases fine grains measuring of ~2 μm in size. In these samples, numerous intergranular pores are present, characterized by relatively modest dimensions, typically not exceeding 2 μm. Occasional patches of a white film, attributed to the presence of additional NiO phase extractions, are observed in these ceramics, primarily near intergranular boundaries, occasionally occupying and partially filling pores.

Moving on to sample 2, the samples reveal larger grains with sizes ranging near 6 μm, with some even reaching dimensions of 9-10 μm. The presence of the white NiO film in these ceramics is primarily confined to the regions adjacent to grain boundaries.

In the case of sample No 3, there is a gradual transformation in the grain structure. The chip structure of these ceramics takes on a more monolithic character, with only isolated pores measuring near 2 μm in size. A distinctive feature is the presence of a bright 10-11 μm thick layer of NiO film on the grain surfaces of these samples.

Sample No 4 exhibits a fully merging grain structure, where only a few individual pores of relatively larger sizes, approximately 4 μm, are observed. In this scenario, the NiO phase forms a uniform layer that spans the entire surface of the ceramics. Notably, the observed distribution of the additional NiO phase within the ceramics bulk is non-uniform, with a more pronounced presence near intergranular boundaries. These phase extractions act as specific trapping centers for positrons that permeate the ceramics.

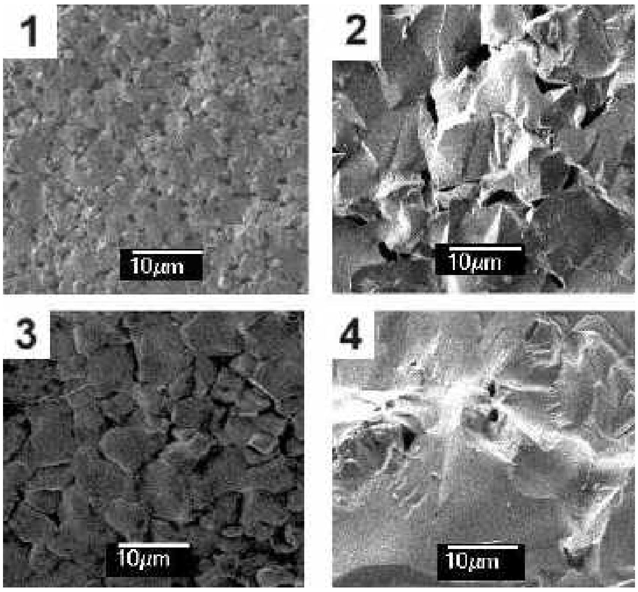
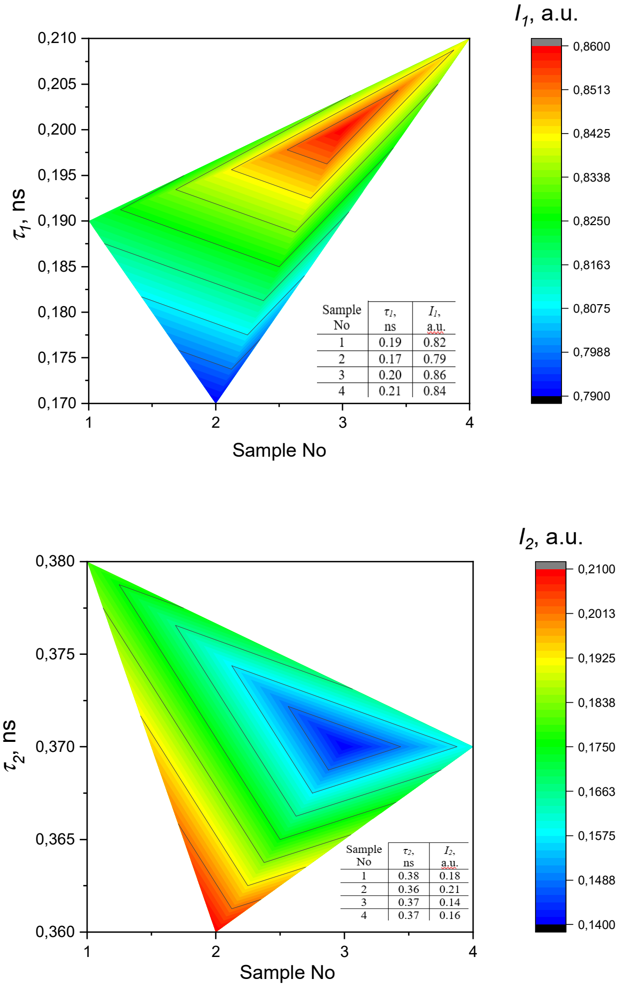


Fig. 1. Microstructural image of $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$ ceramic chipping.

Utilizing a two-state model for positron trapping [16,17], the first spectral component is associated with the primary spinel structure, whereas the second component is attributed to extended free-volume defects situated near intergranular boundaries in close proximity to the additional extracted phases. The intensity I_1 is indicative of the quantity of the primary spinel phase, while the I_2 intensity mirrors the presence of the additional NiO phase near grain boundaries.

The lifetimes τ_1 and τ_2 observed in the sample No 1 serve as representative examples of manganite ceramics $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$, measuring 0.19 and 0.38 ns, respectively (refer to Fig. 2). The influence of an additional phase, comprising 1% NiO, on the positron trapping process is most vividly illustrated by the positron capture rate κ_d, which registers at 0.48 ns^{-1} (see Fig. 3).

Notably, the ceramics in sample No 2 exhibit a distinctive production process, involving an 8-hour annealing at the sintering temperature of monophasic ceramics, reaching 920 °C. The presence of an additional phase, comprising 8 % NiO, is primarily concentrated along grain boundaries. This phenomenon arises from the deliberate reduction of the sintering temperature during the final synthesis stage, lowered from 1200 °C to a lower temperature at a rate of 100°C per hour. As a result, the lifetime of the first component, τ_1, decreases to 0.17 ns, while the intensity of the second component, I_2, experiences a slight rise along with τ_2 (refer to Fig. 1). These adjustments in the fitting parameters correspond to a significant increase in the positron capture rate κ_d, surging to 0.62 ns^{-1}, representing an almost 30% increase. The changes in the parameters of the second component indicate the initial fragmentation of voids with their subsequent agglomeration as the content of the additional phase increases.



Sample No	τ_1, ns	I_1, a.u.
1	0.19	0.82
2	0.17	0.79
3	0.20	0.86
4	0.21	0.84

Sample No	τ_2, ns	I_2, a.u.
1	0.38	0.18
2	0.36	0.21
3	0.37	0.14
4	0.37	0.16

Fig. 2. Lifetimes and intensities of the first and the second components.

In sample No 3, there is a substantial increase in the presence of the additional phase NiO, reaching a concentration of 10 %. This significant rise in NiO content has a detrimental effect on the integrity of the spinel structure, resulting in a noticeable increase in the lifetime parameter τ_1, which now stands at 0.20 ns. Simultaneously, the rate at which positrons are captured, as represented by the positron capture rate, experiences a decrease, now measuring 0.34 ns^{-1}. Continuing this trend, when we further increase the NiO content, pushing it from 10 to 12%, we observe a heightened level of merging within the ceramic structure. This intensified merging process is associated with the increased concentration of NiO. As a result, the previously mentioned parameters, τ_1 and positron capture rate, remain relatively stable without significant changes. In essence, as the NiO content within the ceramic material increases beyond the 10% threshold, it not only impacts the spinel structure but also triggers more pronounced merging phenomena within the material. This underscores the delicate balance between material composition and structural integrity, revealing the intricate interplay between these factors in the context of positron trapping and ceramics.

This increase is attributed to the amplification of energy thermally transferred to the ceramic, particularly during the elevation of sintering temperature from 1200 to 1300°C. The perfection of the spinel structure diminishes further, characterizing the ceramic as "overbaked" with τ_1 increasing from 0.20 to 0.21 ns. Meanwhile, the process of positron capture by bulk defects remains largely unaltered, with τ_2 and κ_d displaying little change. Substantial modifications in $\tau_{av.}$ and (τ_2-τ_b) are not observed, though certain consistent changes in τ_2/τ_b become apparent (Fig. 3).

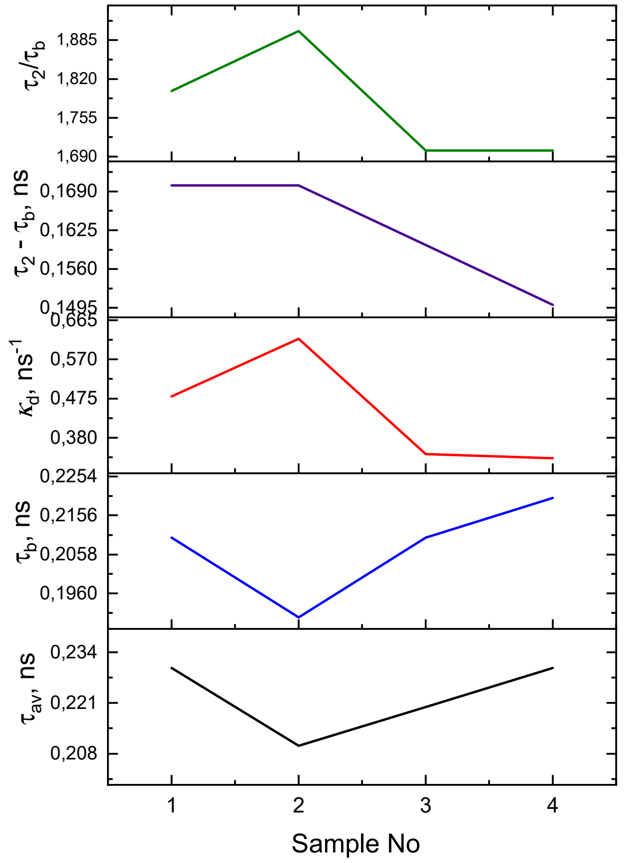


Fig. 3. Positron trapping modes for $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$ ceramic

Notably, the shift in the type of positron capture centers during the monolithization process of ceramics when transitioning from samples of batch No. 2 to No. 3 and from No. 2 to No. 4 is prominently illustrated by a sharp reduction in this parameter from 1.9 to 1.7, signifying a 10 % decrease. It is important to note that in all cases, the nature of positron capture by defects remains consistent, and the size of bulk defects near grain boundaries, estimated from the difference τ_2 - τ_b, corresponds to one or two atomic vacancies [17].

So, the internal nanostructuring of $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$ ceramics evolves contingent upon the content of the additional NiO phase within the material and its presence at grain boundaries. An elevation in NiO content to 8% along grain boundaries triggers an increase in the number of defects (or voids) where positrons are captured, although their sizes undergo some reduction, causing void fragmentation. This, in turn, significantly boosts the positron capture rate.

However, when NiO content is further augmented to 10%, the number of traps where positrons are captured and the parameter κ_d decrease notably. These changes can be attributed to the fact that a substantial portion of the additional NiO phase is no longer concentrated at grain boundaries but rather on the ceramic's surface. The grains expand, forming a monolithic structure and consequently reducing the number of grain boundaries where NiO was initially segregated. Further escalation in the additional phase content to 12% signifies the saturation of the defect formation process, with positron capture parameters exhibiting minimal alterations.

Therefore, concerning the internal volumetric nanostructuring concerning defect and void formation within $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$ ceramics, the optimal batch of samples features a 10% NiO phase content. Alterations in the positron capture rate κ_d most accurately depict the evolution of ceramic structure and the impacts of its nanostructuring and merging, corresponding to differing amounts of thermal energy transferred to the ceramic during sintering [16]. Given the greater presence of grain and pores in the samples of sample No 2, the process of positron trapping within these ceramics intensifies, evidenced by an increase in the positron trapping rate of defects. The component inputs are shown in Fig. 4.

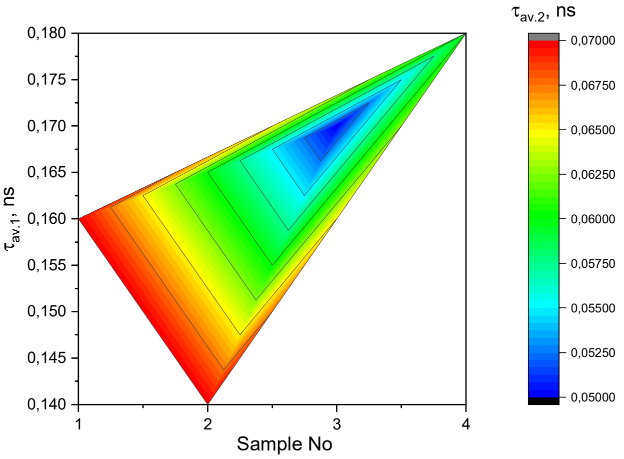


Fig. 4. Inputs of the first and the second components

Conclusion

In conclusion, the results obtained through PAL measurements provide strong evidence of interphase processes within the mixed transition-metal manganite $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$ ceramics. These processes can be attributed to the merging procedures undertaken during technological modification, as well as the presence of an optimal content of additional NiO phase.

It's noteworthy that these PAL findings are highly consistent with the results obtained through microstructural analyses employing X-ray diffractometry and electron microscopy. These complementary techniques further support the notion that structural alterations within the ceramics are a direct consequence of their technological modification. The confluence of evidence from PAL measurements, X-ray diffractometry, and electron microscopy paints a comprehensive picture of the intricate interplay between microstructural changes and technological interventions in these ceramics. This multi-faceted approach provides valuable insights into the underlying mechanisms driving the observed transformations.

Acknowledgment

Authors thank the National Research Foundation of Ukraine for support (project 2020.02/0217 “Light-generation low-dimensional structures with polarized luminescence based on organic and inorganic materials”).

References

[1] S. V. Luniov, (2020). Calculation of band structure of the strained germanium nanofilm, doped with a donor impurity. Physica E: Low-dimensional Systems and Nanostructures, 118, 113954.

[2] S. Luniov, A. Zimych, M. Khvyshchun, M. Yevsiuk, V. Maslyuk, (2018). Specific features of defect formation in the n--Si <P> single crystals at electron irradiation. Eastern-European Journal of Enterprise Technologies, 6(12 (96), 35–42.

[3] S.V. Luniov, Burban, O.V. & Nazarchuk, P.F. (2015). Electron scattering in the Δ1 model of the conduction band of germanium single crystals. Semiconductors 49, 574–578.

[4] L. Liu, Y. Zhou, R. Zheng, M. Gao, P. Zhao, A. Chang, (2023). Low-temperature synthesis and negative temperature coefficient conductivity properties of Mn–Ni–Cu–O spinel ceramics. Journal of Materials Science: Materials in Electronics, 34(5), 371.

[5] V. Seeman, E. Feldbach, T. Kärner, A. Maaroos, N. Mironova-Ulmane, A. I. Popov, A. Lushchik, (2019). Fast-neutron-induced and as-grown structural defects in magnesium aluminate spinel crystals with different stoichiometry. Optical Materials, 91, 42-49.

[6] G. Liu, H. Li, H. Zheng, F. Qian, W. Ma, W. Yang, (2023). Mechanism of in-situ formation of spinel and its effect on the mechanical properties of Al_2O_3–C refractories. Ceramics International, 49(6), 9231-9238.

[7] A. Lushchik, E. Feldbach, E. A. Kotomin, I. Kudryavtseva, V. N. Kuzovkov, A. I. Popov, E. Shablonin, (2020). Distinctive features of diffusion-controlled radiation defect recombination in stoichiometric magnesium aluminate spinel single crystals and transparent polycrystalline ceramics. Scientific reports, 10(1), 7810.

[8] G. Prieditis, E. Feldbach, I. Kudryavtseva, A. I. Popov, E. Shablonin, A. Lushchik, (2019). Luminescence characteristics of magnesium aluminate spinel crystals of different stoichiometry. In IOP Conference Series: Materials Science and Engineering (Vol. 503, No. 1, p. 012021.

[9] C. Teichmann, J. Toepfer, (2022). Sintering and electrical properties of Cu-substituted Zn-Co-Ni-Mn spinel ceramics for NTC thermistors thick films. Journal of the European Ceramic Society, 42(5), 2261-2267.

[10] Y. Liu, W. Deng, X. Chen, Y. Xue, X. Bai, H. Zhang, Y. Xie, (2022). $CaMn_{0.05}Zr_{0.95}O_3$–NiMn2O_4 composite ceramics with tunable electrical properties for high temperature NTC thermistors. Ceramics International, 48(22), 33455-33461.

[11] M. M. Kaci, N. Nasrallah, F. Atmani, M. Kebir, R. Guernanou, A. Soukeur, M. Trari, (2021). Enhanced photocatalytic performance of $CuAl_2O_4$ nanoparticles spinel for dye degradation under visible light. Research on Chemical Intermediates, 47, 3785-3806.

[12] T. Reimann, J. Töpfer, (2021). Low-temperature sintered Ni–Zn–Co–Mn–O spinel oxide ceramics for multilayer NTC thermistors. Journal of Materials Science: Materials in Electronics, 32, 10761-10768.

[13] N. Acharya, R. Sagar, (2021). Structure and electrical properties characterization of $NiMn_2O_4$ NTC ceramics. Inorganic Chemistry Communications, 132, 108856.

[14] H. Dai, Q. Shen, J. Chen, Z. Li, Z. Chen, L. Xie, T. Li, (2023). Improving the magnetic and dielectric properties of $Cu_{1-x}HoxFeO_2$ nanoceramics by tuning the vacancy defects and Fe valence state. Ceramics International, 49(10), 16451-16457.

[15] M. Ghasemifard, M. Ghamari, (2022). Probing the influence of temperature on defects in oxy-hydroxide ceramics by positron annihilation lifetime and coincidence Doppler broadening spectroscopy. Applied Physics A, 128(3), 180.

[16] H. Klym, I. Hadzaman, Y. Kostiv, S. Yatsyshyn, B. Stadnyk, Free-volume defects/nanopores conversion of temperature-sensitive $Cu_{0.1}Ni_{0.8}Co_{0.2}Mn_{1.9}O_4$ ceramics caused by addition phase and monolithization process. Applied Nanoscience, 2022, 12(4), pp. 1347–1354.

[17] H. Klym, I. Karbovnyk, S. Piskunov, A.I. Popov, (2021). Positron annihilation lifetime spectroscopy insight on free volume conversion of nanostructured $MgAl_2O_4$ ceramics. Nanomaterials, 11(12), 3373.

INDEX OF AUTHORS