ISSN 2710-0766
DOI 10.31891/CSIT

THE INTERNATIONAL SCIENTIFIC JOURNAL

COMPUTER SYSTEMS AND INFORMATION TECHNOLOGIES

No 1-2023

МІЖНАРОДНИЙ НАУКОВИЙ ЖУРНАЛ

КОМП'ЮТЕРНІ СИСТЕМИ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

2023

COMPUTER SYSTEMS AND INFORMATION TECHNOLOGIES

INTERNATIONAL SCIENTIFIC JOURNAL

Published since 2020 year *Four time a year*

Khmelnitskyi, 2023, № 1 (10)

Establishers: **Khmelnytskyi National University** (Ukraine)
Associated establisher: **Institute of Information Technologies** (Slovakia)
National Library og Ukraine named after V.I.Vernadsky http://nbuv.gov.ua/j-tit/csit

The journal is included in scientometric databases:
Index Copernicus https://journals.indexcopernicus.com/search/details?id=69998&lang=en
Google Scholar https://scholar.google.com.ua/citations?hl=uk&user=HW1XpMsAAAAJ
CrossRef http://doi.org/10.31891/CSIT

Editors **Hovorushchenko T.,** Doctor of engineering sciences, Professor, Head of the Department of Computer Engineering and Information Systems Khmelnytskyi National University (Ukraine)

Head editorial board **Savenko, O.,** Doctor of engineering sciences, Professor of the Department of Computer Engineering and Information Systems, Dean of the Faculty of Information Technologies, Khmelnytskyi National University (Ukraine)

Executive secretary **Lysenko S.,** Doctor of engineering sciences, Professor of the Department of Computer Engineering and Information Systems Department, Khmelnytskyi National University (Ukraine)

Editorial board:

Hovorushchenko T., Doctor of engineering sciences (Khmelnytskyi, Ukraine), **Savenko O.,** Doctor of engineering sciences (Khmelnytskyi, Ukraine), **Barmak O.,** Doctor of engineering science (Khmelnytskyi, Ukraine), **Lysenko S.,** Doctor of engineering sciences (Khmelnytskyi, Ukraine), **Peter Popov,** PhD (London, Great Britain), **Piotr Gaj,** Dr hab inż (Gliwice, Poland), **Anatolii Gorbenko**, DrSc, Professor (Leeds, Great Britain), **Andrzej Kotyra,** Dr hab inż, Professor (Lublin, Poland), **Andrzej Kwiecień,** Dr hab inż, Professor (Gliwice, Poland), **George Markowsky,** PhD in Mathematics, Professor of Computer Science (Missouri, USA), **Sergii Babichev**, DrSc, Professor (Ústí nad Labem, Czech Republic), **Krak Iu**., Doctor of mathematics and physics sciences (Kyiv, Ukraine), **Yatskiv V.,** Doctor of engineering sciences (Ternopil, Ukraine), **Pastukh O.,** Doctor of engineering sciences (Ternopil, Ukraine), **Romankevich V.**, Doctor of engineering sciences (Kyiv, Ukraine), **Sachenko A.,** Doctor of engineering sciences (Ternopil, Ukraine), **Korobchynskyi M.,** Doctor of engineering sciences (Kyiv, Ukraine), **Bisikalo O.,** Doctor of engineering sciences (Vinnytsia, Ukraine), **Maevsky D.,** Doctor of engineering sciences (Odesa, Ukraine), **Zharikova M.,** Doctor of engineering sciences (Kherson, Ukraine), **Sherstjuk V.,** Doctor of engineering sciences (Kherson, Ukraine), **Berezsky O.,** Doctor of engineering sciences (Ternopil, Ukraine), **Yakovyna V.**, Doctor of engineering sciences (Lviv, Ukraine), **Lupenko S.,** Doctor of engineering sciences (Ternopil, Ukraine), **Shilo G.,** Doctor of engineering sciences (Zaporijjya, Ukraine), **Bobrovnikova K.,** PhD (Khmelnytskyi, Ukraine), **Nicheporuk A.,** PhD (Khmelnytskyi, Ukraine), **Hnatchuk Y.,** PhD (Khmelnytskyi, Ukraine), **Medzatyi D.**, PhD (Khmelnytskyi, Ukraine), **Perepelitsyn A.,** PhD (Kharkiv, Ukraine), **Illiashenko O.,** PhD (Kharkiv, Ukraine), **Izonin I.,** PhD (Lviv, Ukraine), **Horiashchenko S.,** PhD (Khmelnytskyi, Ukraine), **Boyarchuk A.,** PhD (Kharkiv, Ukraine), **Pavlova O.,** PhD (Khmelnytskyi, Ukraine)

Technical editor **Kravchyk Yu,** PhD.

Recommended for publication by the decision of the Academic Council of Khmelnytskyi National University, protocol № 9 from 30.03.2023

Editorial board address: International scientific journal "Computer Systems and Information Technologies", Khmelnytskyi National University, Institutska str. 11, Khmelnytskyi, 29016, Ukraine

☎ **(0382) 67-51-08**

e-mail: csit.khnu@gmail.com

web: http://csitjournal.khmnu.edu.ua/
http://lib.khnu.km.ua/csit_khnu.htm

Registered by the Ministry of Justice of Ukraine
Certificate of state registration of the print media
Series KB № 24924-14864PR dated 12.07.2021

© Khmelnytskyi National University, 2023
© Editorial board "Computer Systems and Information Technologies", 2023

КОМП'ЮТЕРНІ СИСТЕМИ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

МІЖНАРОДНИЙ НАУКОВИЙ ЖУРНАЛ

Засновано в 2020 р. ***Виходить 4 рази на рік***

Хмельницький, 2023, № 1 (10)

Засновник і видавець: **Хмельницький національний університет** (Україна)
Асоційований співзасновник: **Інститут інформаційних технологій** (Словаччина)

Наукова бібліотека України ім. В.І. Вернадського http://nbuv.gov.ua/j-tit/csit

Журнал включено до наукометричних баз:

Index Copernicus https://journals.indexcopernicus.com/search/details?id=69998&lang=en
Google Scholar https://scholar.google.com.ua/citations?hl=uk&user=HW1XpMsAAAAJ
CrossRef http://doi.org/10.31891/CSIT

Головний редактор **Говорущенко Т. О.,** д. т. н., професор, завідувач кафедри комп'ютерної інженерії та інформаційних систем Хмельницького національного університету

Заступник головного редактора. Голова редакційної колегії **Савенко О. С.,** д. т. н., професор, професор кафедри комп'ютерної інженерії та інформаційних систем, декан факультету інформаційних технологій Хмельницького національного університету

Відповідальний секретар **Лисенко С. М.,** д. т. н., професор, професор кафедри комп'ютерної інженерії та інформаційних систем Хмельницького національного університету

Ч л е н и р е д к о л е г і ї

Говорущенко Т.О., д. т. н., **Савенко О.С.,** д. т. н., **Бармак О. В.** д. т. н., **Лисенко С.М.** д. т. н., **Пітер Попов,** доктор філософії (Лондон, Велика Британія), **Пьотр Гай,** д. т. н. (Глівіце, Польща), **Анатолій Горбенко**, д. т. н. (Лідс, Велика Британія), **Анджей Котира,** д .т. н. (Люблін, Польща), **Анджей Квецен,** д. т. н. (Глівіце, Польща), **Джордж Марковський,** к. ф-м. н. (Міссурі, США), **Сергій Бабічев** (Усті над Лабем, Чехія), **Крак Ю.В.** д. ф.-м. н., **Яцків В. В.** д. т. н., **Пастух О.А.,** д. т. н., **Романкевич В.О.,** д. т. н., **Саченко А.О.,** д. т. н., **Коробчинський М.В.,** д. т. н., **Бісікало О.В.,** д. т. н., **Маєвський Д.А.,** д. т. н., **Жарікова М.В.,** д. т. н., **Шерстюк В.Г.,** д. т. н., **Березький О.М.,** д. т. н., **Яковина В.С.,** д. т. н., **Лупенко С.А.,** д. т. н., **Шило Г.М.,** д. т. н., **Бобровнікова К.Ю.,** к. т. н., **Нічепорук А.О.,** к. т. н., **Гнатчук Є.Г.,** к. т. н., **Медзатий Д.М.,** к. т. н., **Перепелицин А.Є.,** к. т. н., **Ілляшенко О.О.,** к. т. н., **Ізонін І.В.,** к. т. н., **Горященко С.Л.,** к. т. н., **Боярчук А.В.,** к. т. н., **Павлова О.О.,** д.ф.

Технічний редактор **Кравчик Ю. В.,** к. е. н., доцент

Рекомендовано до друку рішенням Вченої ради Хмельницького національного університету, протокол № 9 від 30.03.2023

Адреса редакції: **Україна, 29016,**
м. Хмельницький, вул. Інститутська, 11,
Хмельницький національний університет
редакція журналу "Комп'ютерні системи та інформаційні технології"

☎ **(0382) 67-51-08**

e-mail: csit.khnu@gmail.com

web: http://csitjournal.khmnu.edu.ua/
http://lib.khnu.km.ua/csit_khnu.htm

Зареєстровано Міністерством юстиції України
Свідоцтво про державну реєстрацію друкованого засобу масової інформації
Серія КВ № 24924-14864ПР від 12 липня 2021 року

© Хмельницький національний університет, 2023
© Редакція журналу "Комп'ютерні системи та інформаційні технології", 2023

CONTENTS

ЗМІСТ

UDC 004.89
https://doi.org/10.31891/csit-2023-1-1

Nataliya BOYKO, Roman KOVALCHUK
Lviv Polytechnic National University

DATA UPDATE ALGORITHMS IN THE MACHINE LEARNING SYSTEM

This paper analyzes methods for operationalizing anomaly detection, data drift detection, as a data validation step in a machine learning system. A pipeline is a set of data processing elements connected in series, where the output of one element is the input of the next. MLOps is a set of practices aimed at reliable and efficient deployment and support of machine learning models in the real world. We proposed a solution with technologies mentioned in the theoretical paper [1] for operationalizing the Data QC pipeline. Also, we propose to build a Data QC pipeline based on MLFlow, a machine learning cycle manager. We chose MLFlow as a skeleton for building our pipelines. The choice springs from the specifics of the task, problems and the need for ready-made solutions to meet our requirements. Specific explanations are mentioned in the paper [1] both for Data Drift and Data QC pipelines. To construct either Data QC or Data Drift pipeline, we need to wrap the defined solution, divided into steps to the MLFlow. The latter will register all artifacts, metrics and parameters. An artifact in a machine learning system is a result of a process in a pipeline. For example, it could be a trained model, an Excel file, or a feature importance image. The paper considers the following stages of the Data QC pipeline: filtering, anomaly detection, reporting, validation, and comparison of new data with historical. The Data Drift detection pipeline. The Data QC and Data Drift detection pipelines are necessary for data validation and processing in the current machine learning life cycle. The task of the Data QC pipeline is to automate the evaluation and validation of new data. The task is especially important for Time-Series systems in real-time. In this paper, we researched the formation of interactive quality reports, and the anomaly and data drift detection approaches for the Time-Series system. We analyzed approaches to implementing such MLOps architecture with data validation step described with Data QC and Data Drift pipelines.

Keywords: Data Drift, Data QC, Anomaly Detection, MLOps, Data Validation, Machine Learning, Time-Series.

.

Наталія БОЙКО, Роман КОВАЛЬЧУК
Національний університет «Львівська політехніка»

АЛГОРИТМИ ОНОВЛЕННЯ ДАНИХ В СИСТЕМІ МАШИННОГО НАВЧАННЯ

У цій роботі було виконано аналіз методів для операціоналізації пошуку аномалій, виявлення дрифту даних та самого DataQC пайплайну як такого. Проаналізовані підходи до аналізу операціоналізації пайплайну та до операціоналізації виявлення дрифту даних. Виявлення аномалій допомагає нам оцінити чистоту і якість наших даних. Важливо, щоб у моделі не було аномальних викидів, оскільки вони заплутують модель. Також важливо мати послідовні дані без змін у розподілі ознак. Було запропоновано рішення з вибраними технологіями для операціоналізації DataQC пайплайну, визначено наступні кроки для подальшого дослідження. Запропоновано для побудови заданого DataQC пайплайну використати та обґрунтовати власне рішення для пошуку аномалій та виявлення дрифту даних через специфіку задачі, проблеми та відстуності готових рішень які б задовольняли наші вимоги. В роботі розглядаються етапи операціоналізація вищезгаданого пайплайну, який виконує етапи: фільтрування, пошуку аномалій, звітування, валідації, та порівняння нових даних з історичними, для існуючої у системі моделі машинного навчання. Описується складність задачі операціоналізації у реальному світі, яка полягає у постійному оновленні даних, необхідності їх опрацювання та подальшому застосуванні у системі машинного навчання. Також доводиться користь від пайплайну, який б автоматично опрацьовував нові дані. В роботі досліджується проблематика, яку слід розглядати як Time-Series проблему, то при формуванні інтерактивних звітів, перевірці даних на валідність, наявність та пошук викидів, аномалій. Це рішення дозволить нам візуалізувати всі кроки, які виконує конвеєр валідації даних, що дасть змогу іншим розробникам переглянути результат його роботи, не знаючи нюансів його реалізації та не витрачаючи зайвого часу. Також пропонована архітектура MLOps дозволяє відстежувати зміни трендів даних та гарантувати, що модель збереже свою прогностичну ефективність з часом.

Ключові слова: дрифт даних, пайплан, аномалії, операціоналізація, препроцесинг, машинне навчання.

Introduction

The complexity of building solid MLOps architecture in the real world is constantly updating, processing data, model monitoring, and the need for further use in the machine learning system. The benefits of an architecture that automatically processes new data are undeniable. The cleaner our data is, the easier it is for the machine learning algorithm to work with it, and the more predictable the result will be. Sticking to the MLOps principles ensures quality work for all its users: Data Scientists, Software Engineers, and DevOps.

This work aims to create a Data Validation step in our ML system by introducing Data QC and Data Validation pipelines. They are a wrapper of the ready-to-go theoretical solution presented in the previous work. In order to wrap a Python script into several separate pipeline blocks that perform specific jobs, such as anomaly detection, data filtering, or report generation, we got to use a lifecycle manager like MLFlow [6].

With MLFlow, we can record metrics from each experiment through a visual interface, compare its parameters, and evaluate its effectiveness. The mentioned Data QC pipeline consists of the following steps:

1. Loading data from the database.
2. Filtering and preprocessing of data.
3. Search for anomalies in the data.
4. Finding the difference between new and past anomalies.

5. Generating an interactive report on new data.
6. Checking new data on Data Drift.
7. Uploading pre-cleaned data to the database.
8. Logging parameters, metrics, and results of the pipeline execution

We used Pandas and PostgreSQL for data loading and processing. For interactivity of filters - integration with Microsoft Sharepoint. For anomaly detection, machine learning, statistical methods and their combination. For Drift Detection, statistical methods based on testing the null hypothesis of equality of two distributions and rule-based methods. For report generation - Jupyter Notebooks and Holoviz Panel. For logging artifacts, reports, parameters and metrics - MlFlow Constants. For organizing the pipeline - MlFlow Runs. Also, the most important point is that this solution should be On-Premise, that is, work not on the cloud but on a dedicated server of the company.

The decision to place the service on a dedicated server is due to the company's security requirements, which is due to the high cost of confidential data. The project's dataset covers most or all of the employee's actions in the company, his reporting, salary changes, managers' feedback and work history. The risk of such data leakage into the public domain is highly undesirable for the company due to reputational losses, which correspond to monetary losses and data confidentiality issues. Also, data leakage is undesirable due to possible legal problems, leading to reputational and monetary costs.

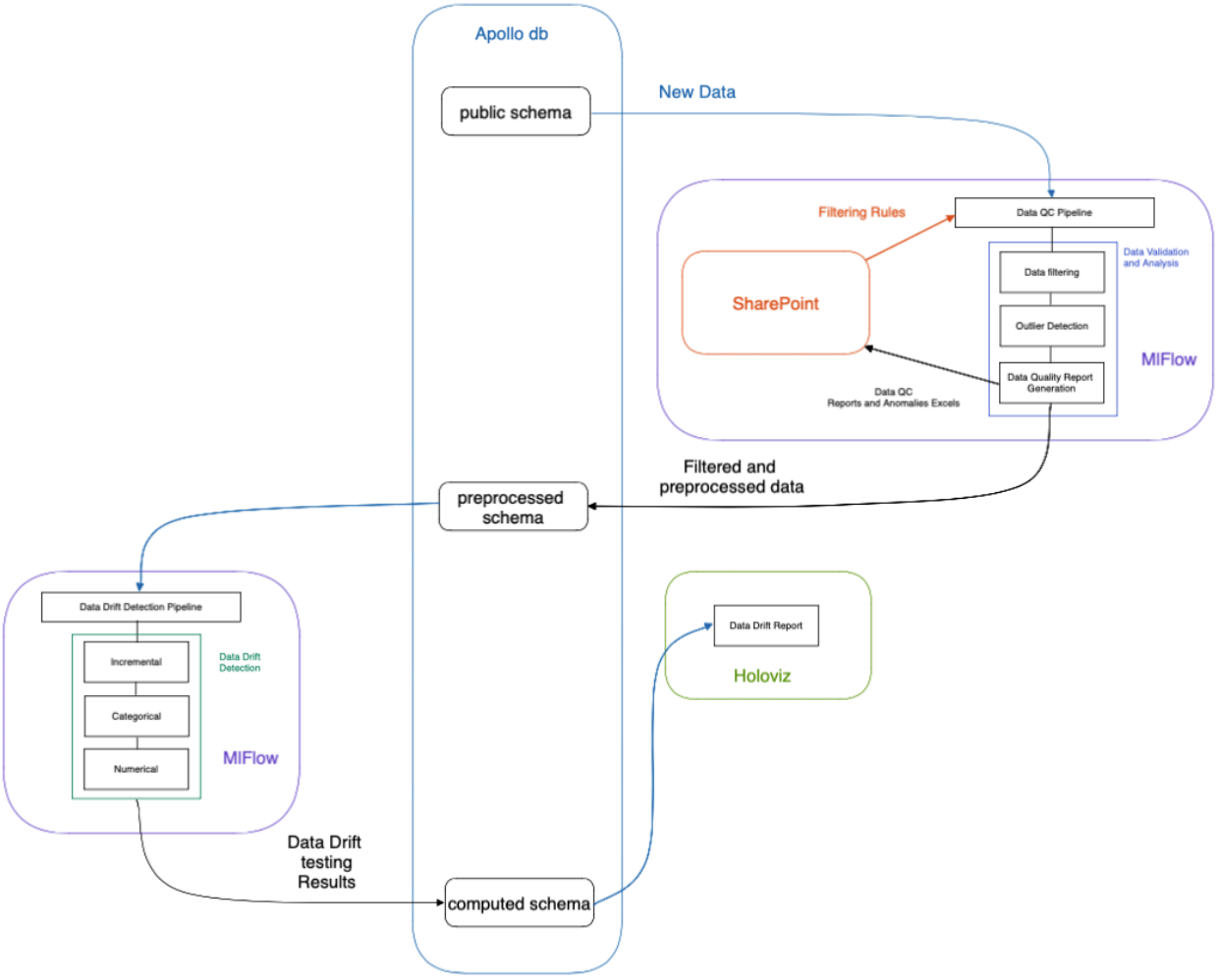


Fig. 1 An example of the architecture of the operationalized solution of DataQC and Data Drift pipelines

According to Fig. 1, a visual representation of the solution's MLOps architecture, we can see that Data Validation is split into DataQC and Data Drift pipelines. Each loads intermediate results into the database. This decision, in addition to architectural rules, for example, separation and isolation of individual tasks into separate pipelines, is also due to practicality. The execution time of the Data Drift pipeline is 7 hours on a dedicated server. In contrast, the Data QC pipeline takes only 15 minutes.

The Data Drift Pipeline is distributed into computing tests and visualizing the result. Firstly, we detect drift in the Data Drift pipeline and write the results to the computed schema of our database. Then, if visualization is necessary, run a Python script that will pull up the necessary data without recalculation. Hosting of visualization page is necessary due to the specifics of working with the Holoviz Panel [3] package.

Although we abstracted ourselves from the model in this work, let us consider it for completeness. We chose a rather complex LightGBM heuristic machine learning model based on decision trees with gradient boosting. The dataset limits the use of transformers or neural networks. Although, with a larger dataset, it could be more efficient in identifying dismissed workers. The following data sources are available:

1. Personal data, e.g. gender, year of birth
2. Status of the employee, e.g. whether he/she is in reserve or dismissed
3. The employee's position, management level and job profile
4. The employee's languages
5. The employee's compensation, bonus history and scheduled salary reviews
6. The customer of the project
7. Project on which the employee is working
8. Certifications that the employee has passed
9. Aggregated Peakon score of the employee on the company and his team
10. Information about the employee's professional review
11. Feedback on the employee from his manager
12. The employee's manager

We can see the depicted model in Fig. 2. Tabular structured data ready for processing is coloured in green. Moreover, purple indicates unstructured text that will be transformed into structured data. This transformation can be performed by the BERT classifier, pre-trained on the GoEmotions [5] dataset, which has a similar specificity to ours, evaluating texts by emotions. Then from these emotions, we can extract the sentiment of the response, for example, whether it is positive or negative. Then, we combine the obtained structured data into single dataset.

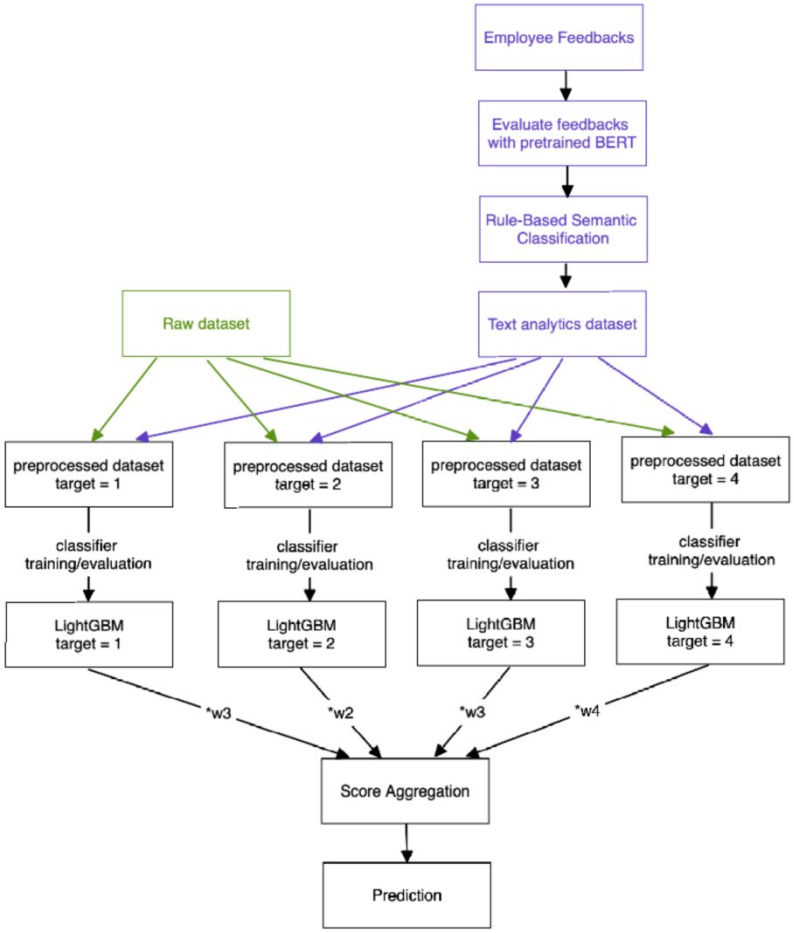


Fig. 2 Example of ensemble architecture of LightGBM models

The "target" is responsible for the hyperparameter of the same name, which corresponds to how much we extend the employee's dismissal into history. How many months in advance do we want to calculate whether the employee will dismiss? For example, with target = 1, we estimate the dismissal in the next month, and with target = 2, in the next two months inclusive.

This work aims to wrap the validation and evaluation of data into pipelines, which in this architecture, in Fig. 2, outputs structured data marked with a green block.

The object of research is a system for predicting the probability of dismissal of a particular employee in a

company after a specified time. A large number of possible independent variables characterizes the system. For example, the model uses about 200 features, some of which are generated.

Since dismissal prediction is a Time-Series task, it is necessary to pay attention to trends and seasonality. This remark also applies to generating interactive reports, verifying data for validity, and anomaly detection.

For instance, there is a trend towards increasing salaries in certain profiles. Hence, the distribution is not stale and is constantly shifting right.

When applying statistical methods, we have to adjust to this error. For instance, assuming Architectors got a plus 10% of their salaries over the next four months, the salary distribution shape for Architectors remains, and we consider this as not an anomaly. Hence, drift and anomaly detection methods should not consider this behaviour anomalous. As an example of an application - search for anomalies in the salary column.

Similarly, we should consider this nuance when checking a column for Data Drift. For instance, the salary distribution for Architectors has dropped by 5% over the last four months. Moreover, the historical data shows us a shifted but identical distribution. Then, the Data Drift should not be detected. Because the salary distributions, except for the conditional mean, are identical. So, we are only interested in detecting the change in the shape of the distribution.

Analysis of recent sources

Analyzing the previous article [7], which describes the theoretical methods for building the data validation step in MLOps architecture, let us briefly recall its components and requirements. So, among the components of the solution of the previous article:

1. Loading data from the database
2. Data filtering and preprocessing
3. Anomaly detection
4. Monitoring the difference between new and past anomalies
5. Generating an interactive report on new data
6. Checking Data Drift on new data
7. Uploading cleaned data to the database
8. Logging parameters, metrics, and results of the pipeline execution.

We have decided to split mentioned in paper [7] data pipeline into two separate ones. Respectively, DataQC and Data Drift Detection pipelines, according to the architecture in Fig. 1 above.

Although the article mentions public packagessuch as Evidently AI [1], we constructed our solution to meet all the requirements. Remembering that we need to wrap these two pipelines in a common architecture, we need to containerize them. So, we will need to split our code in the pipeline into some blocks and steps and use MlFlow to organize logging and tracking experiments and parameters. Tracking is monitoring the results of the execution of a pipeline or a certain job. A pipeline job is the same as a pipeline step. For example, let us single out the Data QC steps of the pipeline from the list above:

1. Loading data from the database
2. Data filtering and preprocessing
3. Anomaly detection
4. Finding the difference between new and past anomalies
5. Generating an interactive report of new data validation
6. Uploading cleaned data to the database
7. Logging of parameters, metrics and results of the pipeline execution

In the list above, for example, from step 3, described in the article [7] of methods and parameters, we form an abstraction in the form of a base and several child classes, one for each implementation of the anomaly detection method, if necessary. Furthermore, one more class would perform all the necessary preprocessing and calls to initiate the work of the previous one, acting as a wrapper. Does building a well-structured OOP code affect the quality of the model? Directly - no, we can copy the linear script of the program, paste MLFlow calls to the API, and finish. However, we will immediately face several anti-patterns in the MLOps world. The first of them is abstraction debt, plain-old-data type smell and glue-code. These anti-patterns are described in detail in the Google study [8], which describes the importance of MLOps architecture in the modern Data Science world. In short, we want to avoid duplicating code but reusing the same logic as much as possible to guarantee the experiment's repeatability. An experiment returning a good model is only possible if we know which parameters to reuse, how to improve the model, or how to debug it. One of the reasons for not reproducible code is just duplicates, where some small error happens. It is considered a good practice to have a well-structured OOP code that is easy to maintain.

After implementing Data QC and Data Drift pipelines with MLFlow, we should containerize these two pipelines using Docker. And since both perform ETL (Extract, Transform, Load), the result of the pipeline execution is always loaded into a common database. By the way, the central idea of our architecture is a shared database with several schemas that display the intermediate results of some pipelines or experiments, see Fig. 1, above. By the way, this necessity arises from banal convenience. For example, we will run the Data QC pipeline, possibly several times with a new batch of data, updated once a month. The result of the Data QC pipeline affects the quality of the model,

as the amount and value of data changes, due to the search and correction of anomalies. We can experiment with the parameters of certain methods mentioned in the article [7]. However, the Data Drift pipeline does not directly affect the model but the drift results. Therefore, there is no need to run it more often than the data is updated. Actually, it was invented for this purpose. For the initial validation of data, when they are updated.

In this article, we will elaborate on the application of the Data Drift pipeline in much more detail than its construction and tests described in the paper [7]. It is much more often used in a full-fledged MLOps architecture. It is especially often when compering with, for example, anomaly detection, which can be found in other articles as a step of a training pipeline or script without recording intermediate results.

Presenting main material

To study the construction of the DataQC pipeline, we identified the necessary features for which we needed to make validation.

Almost every column presented a slightly different approach to solving the problem. For example, we need to apply ANOVA and Percentile filtering to search for anomalies in the multimodal column WageGross. For other numerical columns, we can use the Median + IQR method, for which the only condition is the normality of the distribution. Having evaluated these data types, we formed a list of the necessary anomaly search functions for each of the columns. It is essential to record all the metrics we have obtained. For example, how many employees have null-salary, how many salary values we have filled with past values, the number of anomalies by column, and the parameters we have used. In addition to banal convenience, notation and reproducibility of experiments are key features in building a high-quality artificial intelligence system. For example, after launching the MlFlow client and building a Data QC pipeline, we can observe the parameters with which we ran a certain experiment and what metrics it gave. Anomaly detection or data validation is not a supervised learning task. So, we cannot immediately assess the quality of our Data QC pipeline and how well we handle anomalies. However, we can assess the quality of anomaly processing and data visualization and, ultimately, by assessing the model itself, which is already a supervised-learning task.

In the study on building a given Data QC pipeline, we used and justified our solution for finding anomalies and detecting data drift due to the specifics of the task. Therefore, we should consider the nuances of the MLOps part in more detail.

Among the main requirements: are autonomy, the flexibility of visualization, the flexibility of modifying the logic of the anomaly detection method, and resistance to shifts in distributions.

First, let us consider the requirements for the anomaly search part of the pipeline:

- Search for anomalies by specified columns: APM, WageGross, OnSite, MonthOnPosition and VacancyHistory table.
- Deleting or filling these anomalies from the dataframe.
- Writing these anomalies to Excel files and uploading them to Microsoft SharePoint for automatic monitoring and their elimination at the level of data owners.
- Interactive visualization of the found anomalies.

Since we already have the implementation of the anomaly detection method, the task of operationalizing this step will be to visualize and monitor the result. Considering that we have chosen Bokeh, HVPlot and HoloViews to visualize the found anomalies, we need an interface to display these plots. In this case, two options are available. The first is to group the graph output functions into Jupyter Notebook, run it and convert it to HTML (since the graphs are interactive). And the second is to group the graph output functions into Holoviz Panel. This visualization hosting tool integrates well with the HvPlot platform, HoloViews.

Holoviz Panel is harder to implement as we deploy and describe a service with visualizations. However, the finished service is easy to use. It is always available and is very flexible to modify as we build it.

Converting Jupyter Notebook into an HTML page is much easier but more limited in terms of interactivity, as all the code will be translated into JavaScript. However, this option will display the same charts without interactivity between them, but only in a separate chart.

Fig 3. Example of report visualization using Jupyter Notebooks

For the task of visualizing anomalies, Jupyter Notebook is enough for us, and there is no need to describe the page with Holoviz Panel, although the latter has its advantages mentioned above.

After generating the report, the next step is to convert it to HTML and upload it to Microsoft SharePoint for access by stakeholders and other team members.

Analysis of approaches to bring data drift detection into our MLOps architecture

With data drift detection, we also decided to have our solution. However, we should also mention the cases when we would not use a ready-made package for drift detection. So, let us compare ready-made solutions for data drift detection.

From Fig. 4, consider TensorFlow Data Validation (TFDV) [2]. This tool is an addition to the TensorFlow package and neural networks, part of the ML infrastructure supported by Google - TensorFlow Extended (TFX). This option is unsuitable for us because our model is not a neural network but a LightGBM model. The Whylabs tool does not support non-cloud solutions and, therefore, does not suit us since our MLOps architecture is on-premise. Great Expectations package is also unsuitable because it does not support data drift detection. Evidently AI, already mentioned in the previous paper, was not chosen because it does not support the modification of reports and is limited in modifying testing methods to detect data drift. Due to the above reasons, we have chosen our solution again.

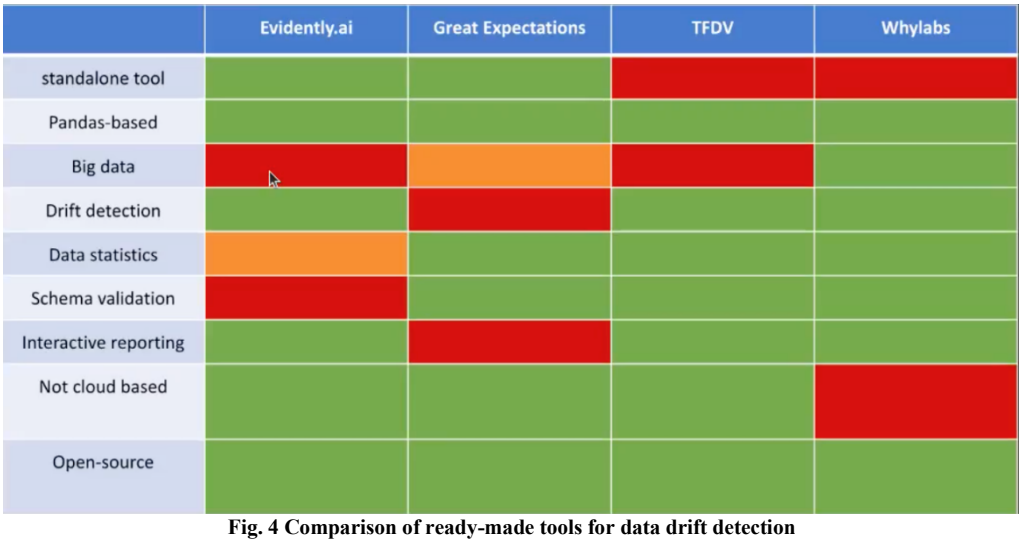


Fig. 4 Comparison of ready-made tools for data drift detection

Since data drift is a more ordinary task, the detection will determine whether the training will occur. We need a visualization tool to interactively compare the graphs of many features to identify where exactly the data drift occurred or what data is invalid. On the contrary, data visualization is more than a fait accompli, which we have to show and not analyze in detail.

Given the previous paragraph, it would be logical to choose the above-mentioned Holoviz Panel tool. Because with it, we can create an interactive page for each available data type, numeric, categorical and incremental, to check the data validity and the presence of data drift.

Analysis of machine learning lifecycle managers

Taking into account the simplicity and linearity of our pipeline, among the available ready-made solutions, for example, TensorFlow Extended (TFX), which we reject due to the lack of TensorFlow and neural networks in our solution. Amazon Sagemaker [4] and similar cloud solutions, which we also reject due to on-premise, we remain on a simpler solution - MlFlow.

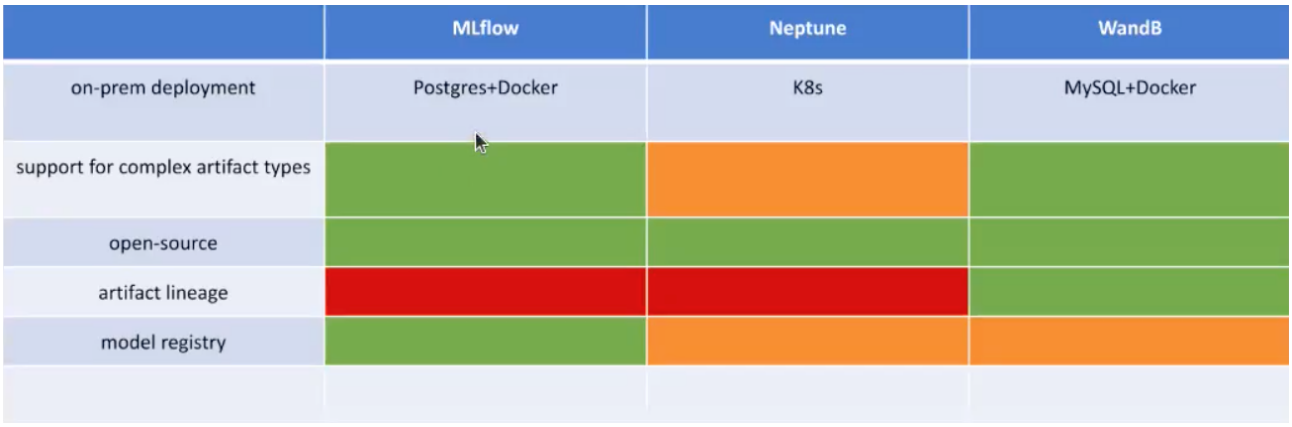


Fig. 5. Comparison of machine learning-lifecycle manaers

We compared machine learning lifecycle managers in Fig. 5.

Note that we chose Docker and Docker-Compose as a tool for deployment and containerization. We do not plan to deploy our solution to any of the clouds since our solution is an On-Premise solution and should be run on our dedicated server. However, when it comes to cloud solutions, Kubernetes (K8s) and Docker are the favourites because of their easy integration and support. It is much more difficult to raise, configure and maintain a K8s cluster on our server than on the cloud. We are responsible for load balancing and expanding the machine's capacity. Moreover, cloud providers usually take this role on themselves.

The next factor for choosing simpler containerization with Docker is that it is much easier to work with and configure. At the same time, Kubernetes focuses on a heavier infrastructure, which includes CI/CD integration. We have this opportunity limited due to the company's security policy.

Conclusions

This paper analyzes methods for efficient deployment and the use of anomaly detection and data drift detection methods in the real world.

We proposed a solution with selected technologies for operationalizing the data validation step of a machine learning system. We identified the following steps for further research, namely:

1. To implement and document the Data QC architecture of the pipeline as a step of data validation before data processing.
2. Operationalize anomaly detection and data drift detection steps using Jupyter Notebook, MlFlow Tracking, Holoviz Panel and Docker.
3. Implement recording of all pipeline artifacts and recording of the filtered dataframe as the final step of ETL (Extract, Transform, Load) of the pipeline.
4. Automatically use the filtered dataframe in the model, regardless of the DataQC of the pipeline.

Among the steps taken to accomplish this work:

1. Anomaly and Data Drift detection in DataQC and Data Drift pipelines, respectively
2. Pipeline management using MlFlow Tracking

Anomaly detection helps us to assess the cleanliness and quality of our data. Let us consider from the point of view of the application of these data. The principle is that the cleaner and better our data, the better our prediction.

It is important for the model not to have anomalous outliers because it confuses the model. Also, it is important to have consistent data without feature distribution changes.

For instance, if the model has learned some feature corresponds to certain qualities, and the relationship between the feature and its meaning changes, the model can not conclude what is happening. We have to catch such cases and automate their detection, which is the second point in the research list above.

Another example is catching an anomaly during model training. Let us imagine feature distribution, and if one has an anomalous record, it shifts and changes the shape of the original data. The model can no longer understand data limits because of a few anomalous records that bring no information gain.

Machine learning management in MlFlow helps us to keep track of the results of experiments and always be sure how our actions have influenced the experiment result. We can always empirically and repeatedly derive any saved experiment and either repeat it or refine it. Also, the tool provides a good visual representation of metrics, which cannot be a disadvantage. We can also save the model itself in MlFlow, which we can reuse in another experiment or automate.

This solution will allow us to visualize all the steps performed by the data validation pipeline, allowing other developers to view the result of its work without knowing the nuances of its implementation and without wasting extra time. We unified the solution with MlFlow and Docker.

Also, our MLOps architecture allows us to keep track of data trend changes. Consequently, ensure that the model will retain its predictive efficiency over time.

References

1. What You Need To Know About Telepresence Robots: What They Are and Use Cases // [Electronic resource] OhmniLabs Writer. – 2021. - Access mode: https://ohmnilabs.com/content/what-to-know-about-remote-telepresence-robot/
2. Hancock E. Pattern Recognition // [Electronic resource] Journal Pre-proof, Vol. 123 – 2021 - Access mode: http://csitjournal.khmnu.edu.ua/
3. Hwang S., Wug Oh S., Kim S. J. Single-shot Path Integrated Panoptic Segmentation // [Electronic resource] Computer Vision and Pattern Recognition. – 2020. – Access mode: https://arxiv.org/abs/2012.01632
4. He K., Gkioxari G., Dollar P., Girshick R. Mask R-CNN // [Electronic resource] .- 2018. – pp. 1-17. - Access mode: https://arxiv.org/pdf/1703.06870.pdf
5. Girshick R. Fast R-CNN // [Electronic resource] arXiv e-prints. – 2015. – Access mode: https://arxiv.org/pdf/1504.08083.pdf
6. Min Read J. S. An Introduction to the COCO Dataset // [Electronic resource] Roboflow Blog. - 2020. - P. 17. – Access mode: https://blog.roboflow.com/coco-dataset/
7. Amazon.com: Brookstone Rover 2.0 App-Controlled Wireless Spy Tank: Toys & Games // [Electronic resource] Amazon.com. – 2020. - P. 1. – Access mode: https://www.amazon.com/Brookstone-Rover-App-Controlled-Wireless-Tank/dp/B0093285XK
8. Double Robotics - Telepresence Robot for Telecommuters // [Electronic resource] Double Robotics – 2021. - P. 2. – Access mode: https://www.doublerobotics.com/double2.html
9. Beam // [Electronic resource] Beam. – 2021. - P. 1. – Access mode : https://suitabletech.com/products/beam.
10. Amazon.com: Appbot Riley Home Safety Movable Camera Robot: Camera & Photo // [Electronic resource] Amazon.com – 2021. - P. 1. – Access mode: https://www.amazon.com/Appbot-Riley-Controlled-Movable-Safety/dp/B01LWXF28H.
11. Gandhi R. R-CNN, Fast R-CNN, Faster R-CNN, YOLO — Object Detection Algorithms // [Electronic resource] Toward data sience. – 2018. – Access mode: https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e
12. Redmon J., Divvala S., Girshick R., Farhadi A. You Only Look Once: Unified, Real-Time Object Detection // [Electronic resource] arXiv e-prints. – 2016. – Access mode: https://arxiv.org/pdf/1506.02640v5.pdf
13. Freeze Tensorflow models and serve on web // [Electronic resource] CV-Tricks.com – 2017. - P. 1. – Access mode : https://cv-tricks.com/how-to/freeze-tensorflow-models/ .
14. Shiledarbaxi N. Guide to Panoptic Segmentation +A Semantic + Instance Segmentation Approach // [Electronic resource] Analytics India Magazine – 2021. – Access mode: https://analyticsindiamag.com/guide-to-panoptic-segmentation-a-semantic-instance-segmentation-approach/ .
15. Hui J. SSD object detection: Single Shot MultiBox Detector for real-time processing // [Electronic resource] Medium – 2020. - P. 1. – Access mode: https://jonathan-hui.medium.com/ssd-object-detection-single-shot-multibox-detector-for-real-time-processing-9bd8deac0e06.
16. Choudhury A. Top 8 Algorithms For Object Detection One Must Know // [Electronic resource] Analytics India Magazine – 2020. - P. 1. – Access mode: https://analyticsindiamag.com/top-8-algorithms-for-object-detection/.
17. Hui J. Object detection: speed and accuracy comparison (Faster R-CNN, R-FCN, SSD, FPN, RetinaNet and…) // [Electronic resource] Medium – 2020. P. 1. – Access mode: https://jonathan-hui.medium.com/object-detection-speed-and-accuracy-comparison-faster-r-cnn-r-fcn-ssd-and-yolo-5425656ae359.
18. Evidently AI; Data Drift Report [Official site of EvidentlyAI]. Retrieved from https://docs.evidentlyai.com/reports/data-drift [in English]
19. TensorFlow Data Validation [Official documentation of Tensorflow]. Retrieved from https://www.tensorflow.org/tfx/data_validation/get_started [in English]
20. Holoviz Panel [Official site of Holoviz]. Retrieved from https://panel.holoviz.org [in English]
21. Amazon SageMaker [Official documentation of Amazon SageMaker]. Retrieved from https://docs.aws.amazon.com/sagemaker/index.html
22. GoEmotions [Official documentation of dataset]. Retrieved from https://ai.googleblog.com/2021/10/goemotions-dataset-for-fine-grained.html
23. MlFlow [Official site of MlFlow]. Retrieved from https://www.mlflow.org/docs/latest/tracking.html
24. Boyko N., Kovalchuk R. Anomaly Detection, Data Drift Detection for Time-Series on Dismissal Prediction System – 2022
25. Sculley D., Holt. G, Golovin D., Davydov E., Phillips T., Ebner D., Chaudhary V., Young M., Crespo J., Dennison. D. Hidden Technical Debt in Machine Learning Systems - 2020

Nataliya Boyko **Наталія Бойко**	Ph.D., Associated Professor at the Department of Artificial Intelligence, Lviv Polytechnic National University, Lviv, Ukraine e-mail: Nataliya.I.Boyko@lpnu.ua https://orcid.org/0000-0002-6962-9363	Доцент кафедри системи штучного інтелекту Національного університету "Львівська політехніка"
Roman Kovalchuk **Роман Ковальчук**	Student at the Department of Artificial Intelligence, Lviv Polytechnic National University, Lviv, Ukraine Middle Data Scientist at SoftServe, Data Science, MLOps, Python, Azure, GCP | Microsoft Certified: Azure AI Fundamentals e-mail: roman.kovalchuk.knm.2019@lpnu.ua https://orcid.org/0000-0001-9039-125X	Студент кафедри системи штучного інтелекту Національного університету "Львівська політехніка"

UDC 004.912
https://doi.org/10.31891/csit-2023-1-2

Olesia BARKOVSKA, Dmytro MOHYLEVSKYI,
Yuliia IVANENKO, Dmytro ROSINSKIY
Kharkiv National University of Radio Electronics, Ukraine

WAYS TO DETERMINE THE RANGE OF KEYWORDS IN A FREQUENCY DICTIONARY FOR TEXT CLASSIFICATION

***T**he paper is devoted to the actual problem of classifying textual documents of the collection by characteristic features, which is used for classifying news, reviews, determining the emotional tone of the text, as well as for forming catalogs of scientific, academic and research works. The research objective is to analyze possible ways to determine the keywords of a document (determination of keywords based on the TF-IDF method, use of keywords defined by the author) for their further use in the classification process as a feature vector. To achieve this goal, the following tasks should be solved: develop a research methodology, including determining the number of keywords using the TF-IDF method; identify the author's keywords; analyze the correlation between the author's keywords and the list of keywords based on TF-IDF; analyze the range and percentage of keywords in the frequency dictionary that include the author's keywords.*

The paper proposes an approach for determining the significant words of a document for their further use as a feature vector in the classification process. In the course of the work, the author's keywords were identified, a partial dictionary was built, and the correlation between the author's keywords and the list of ordered words of the frequency dictionary based on the TF method, which also includes the author's keywords, was analyzed. The determination of the range and percentage of significant words allows for further classification of scientific and research papers when forming thematic catalogs even in the absence of a list of author's keywords that can be used for classification. The proposed sequence of actions includes three key stages: the preparatory stage, the stage of determining the frequency weight of terms, and the stage of determining a significant range of frequency vocabulary for further classification. The results show that the use of the entire input range of frequency dictionary words is redundant and leads to a longer classification time.

Keywords: processing, language, vocabulary, frequency, term, keyword, klassification, feature

Олеся БАРКОВСЬКА, Дмитро МОГИЛЕВСЬКИЙ,
Юлія ІВАНЕНКО, Дмитро РОСІНСЬКИЙ
Харківський національний університет радіоелектроніки

ШЛЯХИ ВИЗНАЧЕННЯ ДІАПАЗОНУ КЛЮЧОВИХ СЛІВ ЧАСТОТНОГО СЛОВНИКУ ДЛЯ КЛАСИФІКАЦІЇ ТЕКСТУ

Робота присвячена вирішенню актуальної проблеми класифікації текстових документів колекції за характерними ознаками, що застосовується при класифікації новин, відгуків, визначенні емоційної тональності тексту, а також для формування каталогів наукових, академічних та дослідницьких робіт. Метою дослідження є аналіз можливих способів визначення ключових слів документа (визначення ключових слів на основі методу TF-IDF, використання ключових слів, визначених автором) для їх подальшого використання в процесі класифікації як вектора ознак. Для досягнення поставленої мети необхідно вирішити такі завдання: розробити методику дослідження, включаючи визначення кількості ключових слів за допомогою методу TF-IDF; визначити авторські ключові слова; проаналізувати співвідношення між авторськими ключовими словами та переліком ключових слів на основі TF-IDF; проаналізувати діапазон і відсоток ключових слів у частотному словнику, що включають авторські ключові слова.

У роботі запропоновано підхід для визначення значущих слів документу для подальшого їх використання у процесі класифікації в якості вектору ознак. В ході виконання роботи було визначено авторські ключові слова, побудовано частковий словник та проаналізована кореляція між авторськими ключовими словами та переліком впорядкованих слів частотного словнику на основі методу TF, до якого також входять авторські ключові слова. Визначення діапазону та відсотку значущих слів дозволяє виконувати подальшу класифікацію наукових та дослідницьких робіт при формуванні тематичних каталогів навіть у випадку відсутності переліку авторських ключових слів, які можна використовувати для класифікації. Запропонована послідовність дій включає три ключові етапи: підготовчий етап, етап визначення частотної ваги термінів та етап визначення значущого діапазону частотної лексики для подальшої класифікації. Результати показують, що використання усього вхідного діапазону слів частотного словнику є надлишковим та призводить, як наслідок, до більшого часу класифікації.

Ключові слова: обробка, мова, словниковий запас, частотність, термін, ключове слово, класифікація, ознака

Introduction

Natural Language Processing (NLP) is a common name for a field that stands at the intersection of computer science and linguistics, as well as many other computationally intensive fields [1,2]. The input data is a variety of conversations between people or text documents, i.e. data that is constantly and spontaneously evolving, regardless of functional style [3] , such as:

- conversational (everyday dialog),
- literary and artistic,
- newspaper and journalistic (news) [4],
- scientific,
- business (official business).

The relationship between text and language can be described as: language is a means of transmitting information; information is contained in text (not in language); text is "built" using language, the language system.

Therefore, languages can be reduced to the analysis of textual data, which are constructed by sequentially combining words into phrases, phrases into sentences, and so on (Figure 1).

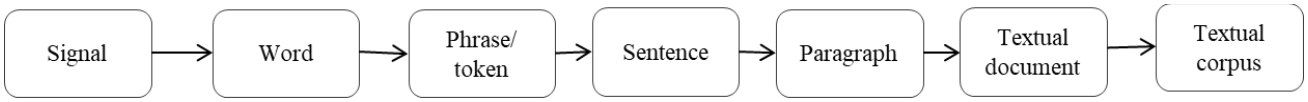


Fig. 1. Hierarchical formation of text data sets

Since computer applications require clear and structured data, natural language processing faces certain challenges and suffers from accuracy. In addition, text analysis methods are highly dependent on language, genre, and topic [5]. Thus, additional customization is always required. That is why many natural language analysis programs include not one but many machine learning models that interact with each other and influence each other [6, 7]. The models can be re-trained on new data, customized for a specific user, and continuously evolve as new information becomes available and various aspects of the program change over time.

Natural language processing is based on solutions to certain common tasks:

- speech recognition [8],
- text classification (by topic, genre, tone, spam filtering) [9],
- information retrieval,
- machine translation [10],
- question and answer systems,
- text and speech generation,
- abstracting the text (annotation, summarization),
- spell checking, etc.

One of the most common and popular tasks in practical applications (Table 1) is the task of classifying text documents [11, 12], which can be defined as the process of assigning a specific category or label to sentences, paragraphs, text reports, or other forms of unstructured text. There are binary (the application task is formulated in terms of yes/no) and multi-class classification (a group of discrete categories) [13].

Table 1

Examples of NLP methods application

Objective	Scope or purpose of the task
Text classification	a stage of the machine translation system; categorize web pages, text documents, and sites into directories; fighting spam (classifying emails, for example); determining the language of the text; displaying the most relevant ads; classification of news by industry; classification of crime types based on incident reports; classification of reviews about goods and services (based on the analysis of the text tone and determination of emotional coloring); monitoring the moral state of the interlocutor (based on analyzing the tone of the text and determining the emotional coloring); determination of hate speech (based on the analysis of the tone of the text and determination of emotional coloring).

Natural language is a very flexible (lexical and structural ambiguity, collocations, idioms) and rich tool for human communication, but its analysis using deterministic rules is a complex process. The flexibility of human interpretation explains why, with only 60,000 symbolic representations (the average vocabulary of a typical university graduate required for communication in a professional context), we are able to outperform computers in instantaneous language understanding. Therefore, similarly fuzzy and flexible computational methods are needed in the software environment. The above, together with the active development of computing technology and the increase in computing power, has led to the active development of computer linguistics methods in recent years and makes the task of natural language processing relevant.

Related works

The well-known classification models include two of the most complex stages:

- preparation of a subject-oriented corpus for building models;
- justification of an analytical solution to a specific applied problem, which traditionally has additional stages: the training stage and the operation stage of the selected model (Figure 2).

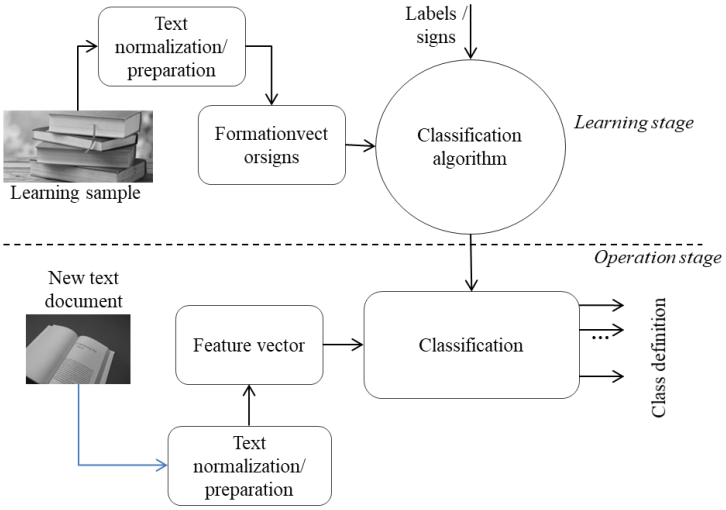


Fig.2. The process of classifying text documents

At the training stage, the document corpus is converted into feature vectors [14, 15]. Then, the document features, along with their labels (the classes that the model is to be trained to recognize), are passed to the classification algorithm, which, based on the comparison of labels and feature vectors, forms its internal state. After the model is trained, the process proceeds to the operation stage, i.e. vectorization of a new text document with the selection of features that are used to perform classification and return the document class label.

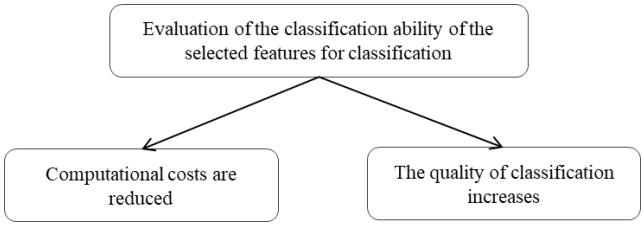


Fig.3. Requirements for the formation of classification features

The classification result, which is measured as classification accuracy (how accurately the model identifies a given class according to the number of times it correctly identified it), classification completeness or sensitivity (how often relevant classes are selected), and F1 score (the weighted average of accuracy and completeness), depends on the generated features that, in turn, are supposed to solve the tasks (Figure 3) [15, 16]:

- reducing computational costs (by reducing the number of features - reducing the space of key terms while maintaining information content and improving interpretability);
- improving the quality of classification (by reducing classification features with low classification ability, for example, by resolving ambiguities in natural language - synonymy, homonymy, polysemy).

The importance of the term for further finding key features and dividing texts into classes is determined by a combination of the following approaches:

- document frequency (DF). It is based on the fact that a significant number of collection terms occur in a small number of documents, and terms with medium or high frequency have the highest information content (subject to preliminary removal of stop words);
- MI (mutual information) - statistical data on the frequency of occurrence of both the phrase as a whole and its words separately;
- IG (information gain) - reaches its maximum when the term is a perfect indicator of the category, i.e., it is present in the document if and only if the document belongs to the class. If the distribution of the term in the category corresponds to the distribution of the term in the collection, then the information gain is 0.

Today, NLP is one of the most researched areas, and many revolutionary developments have been made in this field. NLP relies on advanced computational skills, which is why developers around the world have created many different tools for working with human language. Among such a large number of libraries, some are quite popular and help a lot in performing various NLP tasks, such as NLTK (Natural Language ToolKit), SpaCy, Gensim, SparkNLP, etc.

Let's consider NLP libraries that implement basic text processing methods (Table 2): tokenization, POS tagging, named entity recognizer, dependency parsing, Text Matcher, Chunking, spell checking, Sentiment Detector, pre-trained models, neural network, GPU-based training support [14, 17].

Table 2

Comparative analysis of the advantages and disadvantages of NLP libraries

	Advantages	Disadvantages
Natural Language ToolKit	the most popular and comprehensive NLP library; a lot of third-party extensions; multiple approaches to each NLP task; fast tokenization of sentences; supports the largest number of languages compared to other libraries.	slow; difficult to learn and use; does not support neural network models; there are no integrated word vectors; in sentence tokenization, NLTK only divides the text into sentences without analyzing the semantic structure; it processes strings, which is not very typical for the object-oriented Python language.
SpaCy	the fastest NLP framework; active support and development of the project; easy to learn and use, as it has one optimized tool for each task; handles objects; more object-oriented than other libraries; it uses neural networks to train some models; provides built-in word vectors.	there is not enough flexibility compared to NLTK; sentence tokenization is slower than in NLTK; does not support many languages. There are models for only 7 languages and “multilingual” models.
Gensim	works with large databases and processes data streams; provides tf-idf, word2vec, document2vec vectorization, latent semantic analysis, latent Dirichlet distribution; supports deep learning.	designed for unsupervised text modeling; does not have enough tools to provide a full NLP pipeline, so it should be used with another library (SpaCy or NLTK).
SparkNLP	a complete NLP pipeline; optimized for neural networks; it has a built-in spell checker; expanded with a functional sense detector; scalable.	weak community support.

To summarize, SpaCy and SparkNLP are powerful tools for a complete NLP pipeline and fast processing.

NLTK is also a powerful library that can be used for simple NLP processes or as an NLP pipeline, but it usually has a steeper learning curve.

Gensim is suitable for unsupervised clustering and vectorization NLP.

Proposed technique

Many scientific texts are heterogeneous, due to different functions of influence on the addressee (scientific periodicals and educational literature), as well as the heterogeneity of the subject areas themselves (the language of mathematicians, physicists, linguists, philosophers, etc.).

Scientific texts have different structures and length requirements, depending on the requirements of the publication. An example of a text document that has a scientific style of presentation is the master’s thesis of second-level higher education graduates.

One of the constituent parts of master’s theses are keywords defined by the author, which are one of the features in the classification of text documents.

The considered and analyzed algorithms for selecting features are shown in Table 3.

Table 3

Algorithms for selecting classification features

Name of the algorithm	Characteristics of the algorithm
Frequency of the term	inverse document frequency (TF-IDF) is commonly used to weight each word in a text document according to its uniqueness. Word (token) weights are often used for information retrieval and semantic text analysis. This weight is a statistical measure used to estimate how important a word is to a document in a collection or corpus. In other words, the TF-IDF approach reflects the relevance of words, text documents, and specific categories.
Author’s keywords	is a mandatory component of the master’s thesis abstract. It is determined by the master’s thesis author and makes it possible to establish the main content of the work.

Hence, the research objective is to analyze possible ways to determine the keywords of a document (determination of keywords based on the TF-IDF method, use of keywords defined by the author) for their further use in the classification process as a feature vector.

To achieve this goal, the following tasks should be solved:

- develop a research methodology, including determining the number of keywords using the TF-IDF method;
- identify the author’s keywords;
- analyze the correlation between the author’s keywords and the list of keywords based on TF-IDF;
- analyze the range and percentage of keywords in the frequency dictionary that include the author's keywords.

Experiments

The proposed approach to determining the range of significant words of the frequency dictionary to achieve the goal and solve the identified tasks is shown in Figure 4.

The proposed sequence of actions includes three key stages: the preparatory stage, the stage of determining the frequency weight of terms, and the stage of determining a significant range of frequency vocabulary for further classification.

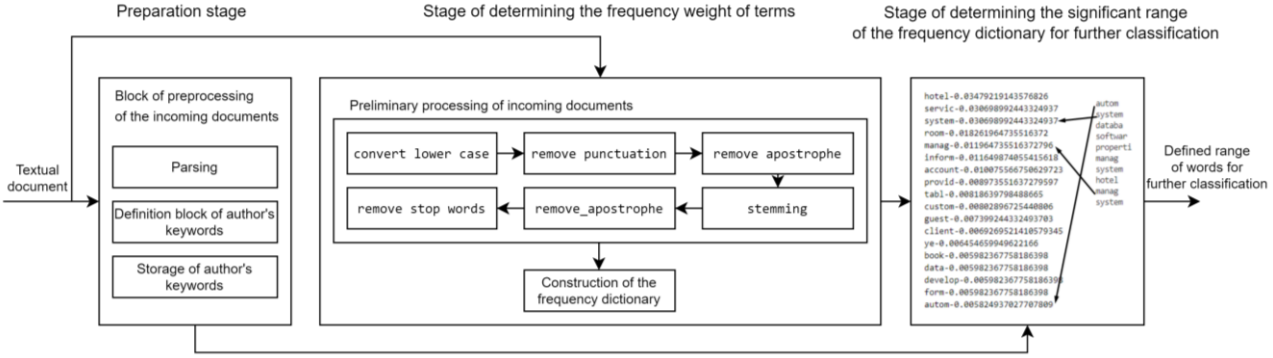


Fig.4. Proposed approach to determining the range of significant words in a frequency dictionary

Let's consider each of the stages separately. The preprocessing stage includes preliminary preparation of the input document, determination and storage of keywords defined by the author. The task of the second stage is to build a frequency dictionary after preprocessing the text document, namely, performing the following steps: converting to a single case, removing punctuation marks, stemming the resulting terms, and removing stop words. The prepared set of terms is the basis for building a frequency dictionary (TF method). The input data for the third stage are the constructed frequency dictionary and the author's keywords. The task of the third stage is to determine the meaningful range by searching all keywords defined by the author in the frequency dictionary.

The results of the proposed approach for three different input documents are shown in Figures 5a, 5b, 5c.

```
Author`s key words  ['autom', 'system', 'databa', 'softwar', 'properti', 'manag', 'system', 'hotel', 'manag', 'system']
0 -  hotel -  0.03479219143576826
2 -  system -  0.030698992443324937
4 -  manag -  0.011964735516372796
17 -  autom -  0.005824937027707809
29 -  databa -  0.004408060453400504
50 -  softwar -  0.0033060453400503777
136 -  properti -  0.0015743073047858943
```

a)

```
Author`s key words  ['chatbot', 'voic', 'assist', 'voic', 'command', 'request', 'analysi', 'system']
18 -  voic -  0.005787781350482315
30 -  chatbot -  0.004777216352779054
37 -  request -  0.004042259990813046
45 -  system -  0.0036747818098300414
59 -  command -  0.0029398254478640333
68 -  assist -  0.002572347266881029
187 -  analysi -  0.0011024345429490124
```

b)

```
Author`s key words  ['imag', 'speedup', 'perform', 'time', 'cpu', 'gpu', 'softwar', 'model', 'hardwar', 'model']
0 -  imag -  0.0317427385892116l7
8 -  gpu -  0.009543568464730291
9 -  time -  0.009128630705394191
15 -  cpu -  0.007468879668049793
18 -  perform -  0.005809128630705394
33 -  hardwar -  0.004149377593360996
93 -  model -  0.0018672199170124482
98 -  softwar -  0.0018672199170124482
234 -  speedup -  0.001037344398340249
```

c)

Figure 5 - Results of the proposed approach for determining the range of significant words in a frequency dictionary based on the identified author's keywords

As can be seen from the results of the application, the partial dictionary is arranged in descending order of word usage. Three documents (document 1, document 2, document 3) in English were used in the experiment.

Determination of the range and percentage of significant words will allow further classification of scientific and research papers when forming thematic catalogs, even in the absence of a list of author's keywords that can be used for classification.

Table 4

Results obtained

	Number of words to be preprocessed, *bt*	Number of words after preproces-sing, *at*	Number of author's keywords, *n*	Author's keywords range in a partial dictionary	Recommended percentage of significant words in the frequency dictionary for further classification	Recommended range of significant words in the frequency dictionary for further classification
Document 1	10154	1296	10	1-137	13,47	1-235
Document 2	8602	1426	8	19-188		
Document 3	7341	1302	10	1-235		

The recommended percentage of significant words for each document containing author's keywords was calculated using the formula below:

$$C = bt/at * 100,$$

where *bt* – the number of words in the input document to be preprocessed; *at* – the number of words in the input document after preprocessing.

The results shown in Table 4 prove that using the entire input range of frequency dictionary words is redundant and leads to longer classification times.

Conclusions

The work is devoted to the urgent problem of classifying textual documents of the collection by characteristic features, which is used to classify news, reviews, determine the emotional tone of the text, and to create catalogs of scientific, academic, and research works.

The paper proposes an approach for identifying significant words of a document for their further use in the classification process as a feature vector.

In the course of the work, author's keywords were identified, a partial dictionary was built, and the correlation between the author's keywords and the list of ordered words of the frequency dictionary based on the TF method, which also includes author's keywords, was analyzed.

Determining the range and percentage of significant words allows further classification of scientific and research papers when forming subject catalogs, even in the absence of a list of author's keywords that can be used for classification. The results show that the use of the entire input range of frequency dictionary words is redundant and leads to a longer classification time.

References

1. Cambria E., White B. "Jumping NLP curves: A review of natural language processing research" //*IEEE Computational intelligence magazine*. – 2014. – T. 9. – №. 2. – C. 48-57.
2. Zhou M. et al. "Progress in neural NLP: modeling, learning, and reasoning" //*Engineering*. – 2020. – T. 6. – №. 3. – C. 275-290.
3. Nadkarni P. M., Ohno-Machado L., Chapman W. W. "Natural language processing: an introduction" //*Journal of the American Medical Informatics Association*. – 2011. – T. 18. – №. 5. – C. 544-551.
4. Al-Jefri M. et al. "Automatic identification of information quality metrics in health news stories" //*Frontiers in public health*. – 2020. – T. 8. – C. 515347.
5. Khan W. et al. "A survey on the state-of-the-art machine learning models in the context of NLP" //*Kuwait journal of Science*. – 2016. – T. 43. – №. 4.
6. Sharir O., Peleg B., Shoham Y. "The cost of training nlp models: A concise overview" *//arXiv preprint arXiv:2004.08900*. – 2020.
7. Gillioz A. et al. "Overview of the Transformer-based Models for NLP Tasks" //*2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. – IEEE, 2020. – C. 179-183.
8. Barkovska O. "Performance study of the text analysis module in the proposed model of automatic speaker's speech annotation" //*Computer systems and information technologies*. – 2022. – №. 4. – C. 13-19.
9. Serdechnyi V. et al. "Model of the internet traffic filtering system to ensure safe web surfing" //*Lecture Notes in Computational Intelligence and Decision Making: Proceedings of the XV International Scientific Conference "Intellectual Systems of Decision Making and Problems of Computational Intelligence"(ISDMCI'2019)*, Ukraine, May 21–25, 2019 15. – Springer International Publishing, 2020. – C. 133-147.
10. Barkovska O. et al. "automatic text translation system for artificial llanguages" //*Computer systems and information technologies*. – 2021. – №. 3. – C. 21-30.
11. Mirończuk M. M., Protasiewicz J. "A recent overview of the state-of-the-art elements of text classification" //*Expert Systems with Applications*. – 2018. – T. 106. – C. 36-54
12. Khan A. et al. "A review of machine learning algorithms for text-documents classification" //*Journal of advances in information technology*. – 2010. – T. 1. – №. 1. – C. 4-20.
13. Jha A., Dave M., Madan S. "Comparison of binary class and multi-class classifier using different data mining classification techniques" //P*roceedings of International Conference on Advancements in Computing & Management (ICACM)*. – 2019.
14. Guo H., Pasunuru R., Bansal M. "An overview of uncertainty calibration for text classification and the role of distillation" //*Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*. – 2021. – C. 289-306.
15. Ogura H., Amano H., Kondo M. "Comparison of metrics for feature selection in imbalanced text classification" //*Expert Systems with Applications*. – 2011. – T. 38. – №. 5. – C. 4978-4989.

16. Lamirel J. C. et al. "A new feature selection and feature contrasting approach based on quality metric: application to efficient classification of complex textual data" //*Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2013 International Workshops: DMApps, DANTH, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD, Australia*, April 14-17, 2013, Revised Selected Papers 17. – Springer Berlin Heidelberg, 2013. – С. 367-378.

17. Jaiswal M. et al. "Detecting spam e-mails using stop word TF-IDF and stemming algorithm with Naïve Bayes classifier on the multicore GPU" //*International Journal of Electrical & Computer Engineering* (2088-8708). – 2021. – Т. 11. – №. 4.

Olesia Barkovska **Олеся Барковська**	Ph.D., Associate Professor of Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine e-mail: olesia.barkovska@nure.ua https://orcid.org/0000-0001-7496-4353	доцент к.т.н., доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна
Dmytro Mohylevskyi **Дмитро Могилевський**	Department of Electronic Computers**,** Kharkiv National University of Radio Electronics, Kharkiv, Ukraine **e-mail:** dmytro.mohylevskyi@nure.ua	магістрант**,** Харківський національний університет радіоелектроніки, Харків, Україна
Yuliia Ivanenko **Юлія Іваненко**	Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine; e-mail: yuliia.nazarenko@nure.ua	магістрант**,** Харківський національний університет радіоелектроніки, Харків, Україна
Dmytro Rosinskiy **Дмитро Росінський**	Senior Lecturer of Department of Electronic Computers, Kharkiv national university of radio electronics, Kharkiv, Ukraine, e-mail: dmytro.rosinskyi@nure.ua	старший викладач кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна

UDC 004.853:007.2
https://doi.org/10.31891/csit-2023-1-3

Sergii BOZHATKIN, Viktoriia GUSEVA-BOZHATKINA,
Tetyana FARIONOVA, Volodymyr BURENKO, Bohdan PASIUK
Admiral Makarov National University of Shipbuilding

EMERGENCY NOTIFICATION COMPUTER SYSTEM VIA TELECOMUNICATION EQUIPMENT OF THE ORGANIZATION'S LOCAL NETWORK

In the event of an emergency, there are still actions that people must take to save themselves. Currently, everyone has a mobile phone. Almost all establishments have an open Wi-Fi network. A model of the system that, when connected to the network, informs about the threats that have arisen and the actions that citizens must take to avoid damage. The alert system works around the clock. It complements the existing fire alarm and security systems. In the course of the work, an analysis of the existing models of cybersecurity threats for warning systems in emergencies was carried out, which showed that the requirements for the civil protection warning system currently need to be modernized. Therefore, the purpose of the work is to design and develop an extended cybersecurity threat model. The key aspects of the cybersecurity threat model are identified. A model of an intruder of such a warning system is presented. An extended cybersecurity threat model has been built using the Cyber Kill Chain.

At this stage of the study, data were obtained that allow us to conclude about the use of the Cyber Kill Chain model. When applied to a typical threat model, the result gives a broader view of the threats to the information system (including actors, typical hacker software, devices that may eventually become hacker tools when the system is hacked).

The use of modeling to study each of the structural components of the warning system is determined to be appropriate. This is justified by the fact that it is impractical to conduct a real experiment, especially with the reproduction of cyber incidents, due to significant financial and labor costs. This approach is also effective when it is necessary to conduct an analysis of the designed system, which does not yet physically exist in this organization.

Keywords: emergency, notification computer system, public wireless access point, alert nodes, organization's local network, cybersecurity, model of cybersecurity threats, Cyber Kill Chain

Сергій БОЖАТКІН, Вікторія ГУСЄВА-БОЖАТКІНА,
Тетяна ФАРІОНОВА, Володимир БУРЕНКО, Богдан ПАСЮК
Національний університет кораблебудування імені адмірала Макарова

КОМП'ЮТЕРНА СИСТЕМА ОПОВІЩЕННЯ ПРО НАДЗВИЧАЙНІ СИТУАЦІЇ ЗА ДОПОМОГОЮ ТЕЛЕКОМУНІКАЦІЙНОГО ОБЛАДНАННЯ ЛОКАЛЬНОЇ МЕРЕЖІ ОРГАНІЗАЦІЇ

У разі надзвичайної ситуації необхідно вжити всіх необхідних заходів, щоб врятувати людей. Майже всі заклади мають відкриту мережу Wi-Fi для співробітників. На теперішній час у кожного з них є мобільний телефон. На основі телекомунікаційного обладнання корпоративної мережі та смартфонів у якості хостів такої мережі можливо побудувати систему оповіщення, яка працює цілодобово. Така системі також може бути доповнена існуючими системами пожежної сигналізації та охорони.

Моделювання системи оповіщення полегшує вивчення поведінки об'єктів з метою покращення функціональності та зменшення вартості такої системи під час її створення, подальшого перетворення і розвитку. До того ж, в такій моделі системи має бути враховано не тільки інформування про загрози, але й дії, які громадяни мають виконати, щоб уникнути небезпеки та мінімізувати збитки на виробництві. У ході роботи також проведено аналіз існуючих моделей загроз кібербезпеці для систем оповіщення про надзвичайні ситуації.

Комплексний підхід до моделювання всіх зазначених складових комп'ютерної системи оповіщення про надзвичайні ситуації за допомогою телекомунікаційного обладнання локальної мережі організації (ЛОМ) показав, що вимоги до системи оповіщення цивільного захисту наразі потребують модернізації. Тому у роботі розглянуті питання проєктування такої системи. Також розроблено розширену модель загроз кібербезпеці системі оповіщення через обладнання ЛОМ. Визначено ключові аспекти моделі загроз кібербезпеці. Представлено модель порушника такої системи оповіщення. Розширену модель загроз кібербезпеці було створено за допомогою Cyber Kill Chain.

Використання моделювання для дослідження кожної зі структурних складових системи оповіщення визначається як доцільне, тому що реальний експеримент, особливо з відтворенням кіберінцидентів, проводити недоцільно через значні фінансові і трудові витрат, а також при необхідності проведення аналізу проектованої системи, яка ще фізично не існує в даній організації.

Ключові слова: надзвичайна ситуація, комп'ютерна система сповіщень, публічна бездротова точка доступу, вузли сповіщень, локальна мережа організації, кібербезпека, модель загроз кібербезпеці, Cyber Kill Chain

Introduction

One of the main ways of protecting the population from emergencies is a timely notification of the danger in the situation that has arisen as a result of its development, as well as informing about the procedure and rules of behavior in the context of the emergency.

Today there is a revision of the requirements regarding modern alert systems (AS), which were created for civil protection tasks using automated systems of centralized notification, communication networks, radio broadcasting. There is a transition to new structures of organization of such systems, taking into account the current

state of technical means of communication, protection against unauthorized access, and distribution of malicious software.

Emergency alert systems serve as a critical link in the chain of crisis communication, and they are essential to minimize loss during emergencies. Acts of terrorism and violence, chemical spills, amber alerts, nuclear facility problems, weather-related emergencies, flu pandemics, and other emergencies all require those responsible such as government officials, building managers, and university administrators to be able to quickly and reliably distribute emergency information to the public [1, 2].

Loudspeaker in such places attracts attention and may provide the necessary information for further action but at the same time the presence on the screen of a smartphone, tablet, laptop clear scheme using microservices, evacuation plan and instructions for actions of the population especially with hearing impairments, which will minimize the time to make decisions about an emergency response or mitigation measures [3].

Therefore, the danger must be notified first via calls, howls, sirens, etc. But currently, everyone has a mobile phone, and almost all establishments have a Wi-Fi network. Consequently, a system that, when connected to the network, informs about the threats that have arisen and the actions that citizens must take to avoid damage, is necessary important at this time.

Modern AS and information support are created to solve the assigned tasks based on automated centralized warning systems, communication networks, and broadcasting [4]. Also, when building notification systems, it is necessary to take into account the security of access points (AP) and system servers from hacking by intruders, preventing DDoS attacks, etc. [5].

The security system of information systems (such like AS) is not built by itself. It is based on threat models and intruder models. The threat model itself is a document that lists and describes possible threats to the information security of the organization/enterprise, the probability of implementation, and the consequences of their action.

The idea of creating a public alert system via Wi-Fi is quite new and interesting for research and implementation. Since we are talking about APs of wireless connection, there is a threat of various kinds of attacks. To determine the cybersecurity of such a system, it is proposed to develop an extended threat model based on the Cyber Kill Chain model and to study the difference with a typical threat model [6, 7].

Research interests – cybersecurity, models of hacker attacks.

The aim of the present study is to increase the accuracy of the threat model on the example of the development of a wireless alert system by developing an extended threat model through detail using the Cyber Kill Chain model.

Objectives of the study.

- Identify typical models on which typical final threat models and violator models are based
- Analyze the type of connection between digital network points
- Identify typical attack methods for this type of connection
- Develop assessment criteria in an extended threat model based on Cyber Kill Chain items
- Develop and present the results of the study of the alarm system via wireless communication

The relevance of the study is based on the fact that the rapid acceleration of the development of computer technology entails more complex and diverse attacks, which are difficult to describe with existing typical models of threats. Therefore, there is a need to develop a more accurate and detailed model with expert assessments for each stage of the security system being tested for hacking.

The object of the study Methods of extending a typical threat model using the Cyber Kill Chain model.

The subject of the study is the properties of the information system cybersecurity violator model after the implementation of key points of the Cyber Kill Chain model.

The relevance of the study is to find new, more detailed, more effective models for building a threat model of information systems. At present, not every business or government organization in Ukraine is thinking about the problems of creating a model of threats to their information systems. Often it is due to the negligence or ignorance of the system administrator that vulnerabilities remain in the systems, which an attacker can exploit at his own discretion without any problems. His actions can result in the complete destruction of information, as well as its theft or sale on the black market.

In this regard, it should be noted that the most effective description tools are the model of intruder and the model of threats to the information system, which simultaneously provides a representation of two key issues: identifying the system actor that can harm the information system and attack vectors.

Existing standard approaches and models are quite general and do not describe at what stage the actor (employee or attacker) can perform intentional or unintentional actions that may harm the information system. This research is based on the development of a wireless warning system for students of the Admiral Makarov National University of Shipbuilding.

Problem statement

The threat itself is a security flaw or omission that can be exploited by attackers. The presence of a threat does not mean the inevitable possible leakage of information: this suggests that attackers have a theoretical possibility of unauthorized access to the personal data of the enterprise.

Like any normative document, the threat model is built on a certain model: the title page, a list of terms, definitions, and abbreviations, content, main part, and appendices [8].

To create a model, it is necessary to analyze the data obtained during the audit of the information system (IS). This will help identify system weaknesses; understand what will threaten it; where the threat may come from and by what means it will be possible to neutralize it or prevent its detection in advance [9].

Sources of threats – a section that also needs to be reflected in the model. These can be external or internal intruders, viruses, or software and hardware bookmarks.

When compiling a threat model, the level of initial security is determined. This is a global parameter that is determined once and does not change depending on the threat. Then the actual threats are highlighted and unnecessary ones are excluded – those that do not harm the system. Threats that have not been ruled out are included in the model with a description.

Threat modeling is still in some ways an art as much as a science, and there is no single canonical threat modeling process. The practice of threat modeling draws from various earlier security practices, most notably the idea of "attack trees" that were developed in the 1990s. In 1999, Microsoft employees Loren Kohnfelder and Praerit Garg circulated a document within the company called "The Threats to Our Products" which is considered by many to be the first definitive description of threat modeling [10].

In [11, 12] present threat models in the form of a list of possible IP vulnerabilities (such as DoS, DDoS, sniffing, packet header substitution, etc.), but objects like an alarm system are subject to much more thorough inspection and the developed model should be to some extent more detailed.

The need for a Wi-Fi network is that the marketing policy of the center provides visitors and employees of the center and shops the opportunity to access the Internet. There are many times when you need to access the Internet not only from your computer or laptop, but also from portable devices that allow you to optimize your workflow at the expense of modern network infrastructures – video conferencing, IP telephony, e-mail, server management, and network devices.

The high level of security of Wi-Fi indicates its advantages when used in public places where information security is one of the main criteria of the network. To protect against unauthorized access to the alert node and save the database of connected subscribers the RADIUS Protocol is the most common AAA (Authentication, Authorization, and Accounting) protocol now developed to transmit information between application programs (Fig. 1).

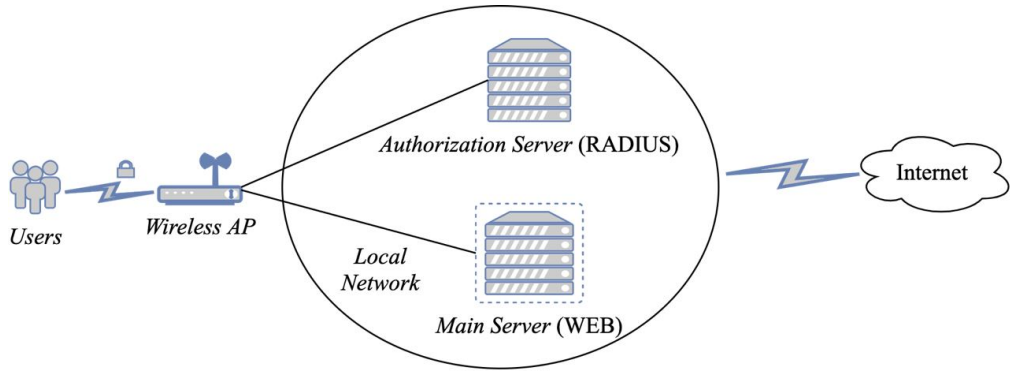


Fig. 1. Alert node scheme

It should be noted that this model can be applied to employees of the *Local Network* of the enterprise, who are granted access to the Wi-Fi network with a RADIUS server (*Authorization Server*) and a notification server (*Main Server*). This means, that the connection between *Wireless AP* and *Users* (employees) has an encrypted connection [13].

To describe a problem, we need to understand what is typical cybersecurity threats exist for such systems. There are many specific types of attacks on Wi-Fi networks [14]:

- Hacking WPA/WPA2 passwords (handshake catching);
- WEP attack;
- Hacking WPS Pin;
- WPA downgrade;
- Replacing a true AP with a fake (for catching login and password to connect to the AP and compromised it);
- Attack on Wi-Fi access points from the global and local networks;
- Denial of Service Attacks (Wi-Fi DoS);
- Attacks on specific services and functions of routers;
- Keyloggers on mobile devices;
- Hijacking;

– Social Engineering.

First, let's define the model of information security violator. For this purpose, the typical model shown in Fig. 2 will approach.

This typical model of cybersecurity violator of the notification system in the enterprise reflects only general information and carries almost no semantic load for a cybersecurity specialist or system administrator. In this case, the next step is to build a threat model for the notification system by using the so-called "CIA Triad": Confidentiality, Integrity, and Availability [15]. Details of the implementation of this approach in different types of activities have shown in Fig. 3.

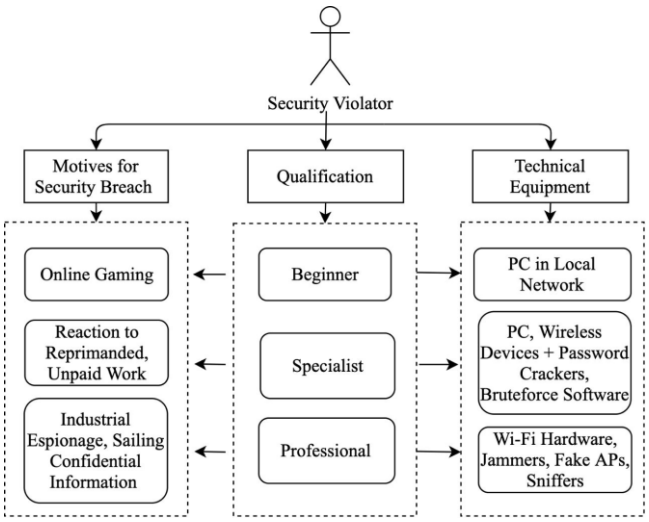


Fig. 2. Typical model of information security violator

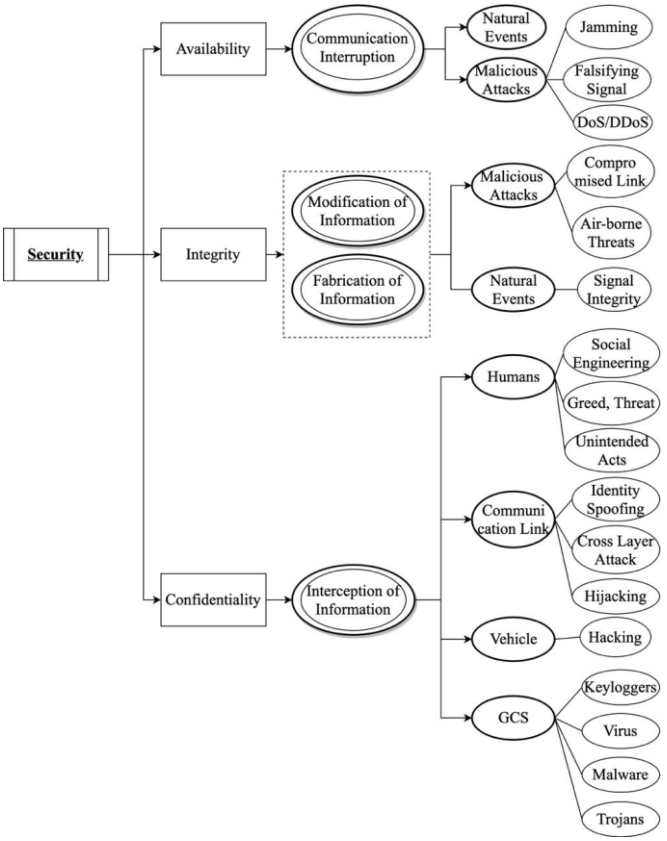


Fig. 3. Typical threat model

The problem is that these models are guided by the reflection of general recommendations that should be considered, but in no way indicate the real possible violators and at what stage they may begin to exploit system vulnerabilities.

Given the above, it is necessary to conduct research and build a threat model that more accurately transmits information about possible vulnerabilities of information systems on the example of a notification system via wireless communication.

Experiment

Threat modeling is a structured process through which IT pros can identify potential security threats and vulnerabilities, quantify the seriousness of each, and prioritize techniques to mitigate the attack and protect IT resources.

This broad definition may just sound like the job description of a cybersecurity professional, but the important thing about a threat model is that it is systematic and structured. Threat modelers walk through a series of concrete steps to fully understand the environment they're trying to secure and identify vulnerabilities and potential attackers.

Within cybersecurity, we see many terms used within military operations, including demilitarized zones (DMZs), defense-in-depth, and APT (Advanced Persistent Threat). Another widely used term is the kill chain where military operations would attack a specific target, and then look to destroy it. A defender will then look to break the kill chain and understand how it might be attacked. An example of the kill chain approach is "F2T2EA", where we Find (a target), Fix (on the location of the target), Track (the movement of the target), Engage (to fix the weapon onto the target), Assess (the damage to the target). A core of this approach is the provision of intelligence around the finding, tracking, and assessment of the target.

One of the most used cybersecurity models to understand threats is the kill chain model and was proposed by Lockheed Martin. Yadav and co-authors determine that the technical nature of the key stages of an attack, include Reconnaissance, Weaponize, Delivery, Exploitation, Installation, Command & Control, and Act on Objective (Fig. 4).

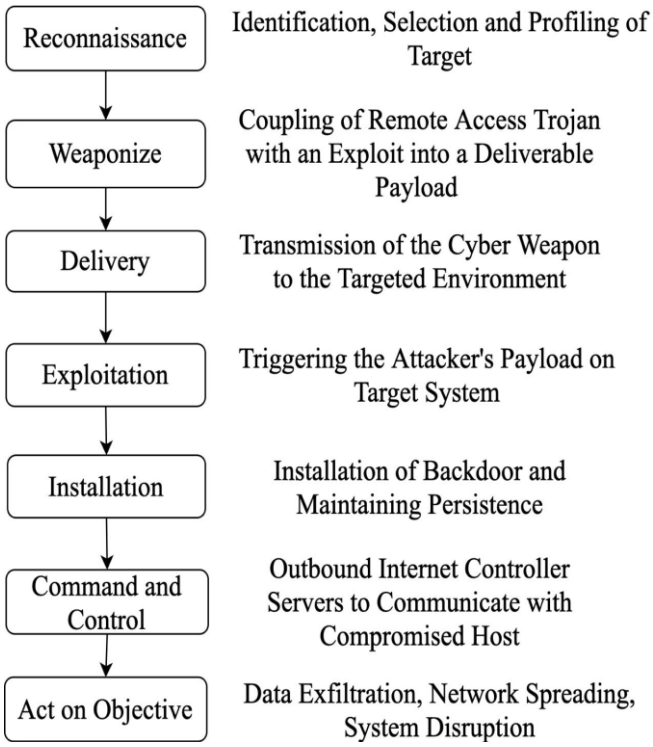


Fig. 4. Simple Cyber Kill Chain model

Each stage is related to a certain type of activity in a cyberattack, regardless of whether it's an internal or external attack.

1) Reconnaissance

The observation stage: attackers typically assess the situation from the outside-in, to identify both targets and tactics for the attack

2) Intrusion

Based on what the attackers discovered in the reconnaissance phase, they're able to get into your systems: often leveraging malware or security vulnerabilities

3) Exploitation

The act of exploiting vulnerabilities, and delivering malicious code onto the system, to get a better foothold

4) Privilege Escalation

Attackers often need more privileges on a system to get access to more data and permissions: for this, they need to escalate their privileges often to an administrator:

1) Lateral Movement.

Once a hacker enters the system, he can move laterally to other systems and accounts to gain more leverage: whether that's higher permissions, more data, or greater access to systems;

2) Obfuscation/Anti-forensics.

To successfully pull off a cyberattack, attackers need to cover their tracks, and in this stage, they often lay false trails, compromise data, and clear logs to confuse and/or slow down any forensics team;

3) Denial of Service.

Disruption of normal access for users and systems, to stop the attack from being monitored, tracked, or blocked;

4) Exfiltration.

The extraction stage: getting data out of the compromised system.

The term "Kill Chain" was originally used as a military concept related to the structure of the attack. The idea is to effectively prevent or counteract the opponent in the various phases of the attack lifecycle.

To attribute cyber threats effectively, it is necessary to identify them based on their attack patterns in different phases of the kill chain. These are tactics, techniques, procedures, and the tools used (software). Tactics are the goals or states that an attacker tries to achieve to complete their mission. A technique is how a specific behavior or activity achieves that goal or state. A tactic can have many techniques and a technique can have many tactics. Procedures and software identify the tools or steps used to complete a series of actions conducted in a certain order or manner. To achieve one of these steps an APT can use many tactics. In turn, these tactics are accomplished by using one or many techniques and/or software tools.

Taking into account the above, we will define the criteria according to which the model of cybersecurity threats of the alarm system via wireless communication will be built based on the following interrelated points.

- stage of the model "Cyber Kill Chain";
- description of threat at this stage;
- typical tools of the attacker at this stage;
- entry points of the attacker at this stage (possible security gaps);
- danger actor (person or department with whose hands you can enter the information system) at this stage;
- the mechanism for implementing malicious actions at this stage;
- the level of danger of the attacker at this stage (scale 1–5, where 1 – the threat of breakage does not entail any consequences; 5 – possible complete failure of the system);
- the current level of implementation of preventive actions at the stage of the model "Cyber Kill Chain".

Given the defined criteria for building an extended model of cyber threats to the notification system via wireless communication, the results are presented in Table 1.

Table 1

Enhanced cybersecurity threat model

Stage	Description of threat	Typical tools	Entry points	Threat actors	Implementing mechanism or device	Level of danger (1–5)	Current level of preventive actions
Reconnaissance	Target investigation	nmap, vuln search, aircrack-ng phishing; sqlmap	SSID is visible, known employee's e-mail; Captive Portal website	Cracker (hacker)	–	2	1**
Weaponize	Making a payload or phishing link, basics on the investigation of a target	Msfvenom, custom payload; compromised e-mail account	Weak Wi-Fi encryption mechanism; handshake capture; admin's mail service; Captive Portal website	Cracker	Wi-Fi access point; e-mail service without antispam; weak SQL database	3	3**
Delivery	Launch exploit	Meta-sploit framework, msfvenom, bash shell	Outside IP address for Internet access; Captive Portal website	Cracker	Server, smartphone or laptop	4	3**
Exploitation	System infection	Malware, malicious connect	Mobile devices with Bluetooth vulnerabilities open ports on AP*	Cracker, administrator	Server (Radius, main)	4	2**

Stage	Description of threat	Typical tools	Entry points	Threat actors	Implementing mechanism or device	Level of danger (1–5)	Current level of preventive actions
Installation	Starting payload's session	Meterpreter, bash shell, malware attack	Vulnerable service, AP, or server	Cracker, administrator	Server or laptop, mobile device	4	3**
Command and control	Using payload's session to take control of the overall system	Meterpreter, bash shell	Payload on a server or vulnerable service exploit	Cracker	Compromised devices in a system	5	4**
Act and objective	Data compromising system disruption	Meterpreter, bash shell, ransom-ware attack	Payload on a server or vulnerable service exploit	Cracker	Compromised devices in a system	5	5**

*Open ports, which are using for Radius server services or main server access
**Average assessment of experts (research will be conducted in the next scientific article)

In cases of the extreme complexity of the problem, its novelty, insufficient information available, the impossibility of mathematical formalization of the solution process, one has to turn to the recommendations of competent specialists who know the problem perfectly – to experts. Their solution to the problem, argumentation, formation of quantitative assessments, processing of the latter by formal methods are called the method of expert assessments.

Expert assessment involves the creation of a collective opinion that has greater capabilities compared to the capabilities of an individual. The source of collective opinion is the search for weak associations and assumptions based on the experience of an individual specialist. The expert approach has great potential for solving problems that cannot be solved in the usual analytical way.

Let's provide expert estimates for each point of the model presented above about the current level of preventive actions (Table 2).

Table 2

The current level of implementation of preventive actions at the stage of the model "Enhanced cybersecurity threat model"

Stage	Expert 1	Expert 2	Expert 3	Average score
Reconnaissance	1	2	1	1
Weaponize	2	3	3	3
Delivery	3	3	3	3
Exploitation	3	2	2	2
Installation	4	3	3	3
Command and control	4	4	4	4
Act and objective	5	5	5	5

Conclusion

At this stage of the study, data were obtained that allow us to conclude about the use of the Cyber Kill Chain model. When applied to a typical threat model, the result gives a broader view of the threats to the information system (including actors, typical hacker software, devices that may eventually become hacker tools when the system is hacked).

It was also proposed to introduce expert assessments to determine the degree of security of the information system at a certain stage of the developed model, which will be studied in more detail and the results will be provided in future research papers.

Further research should be aimed at improving the extended threat model of information systems. In the first stage, the integration of this model should be carried out on a small segment of the network to collect data on the security of the information system and identify possible gaps in its security.

Further analysis and improvement of such a model will show how effective the information system will be in terms of cybersecurity and will help to immediately understand and correct deficiencies. Network security issues deserve special attention, especially in the case of the integration of next-generation networks in such important areas of infrastructure as, for example, wireless alert systems, electricity, or energy delivery system.

Also, new standards for the organization of computer networks can be applied in other areas of telecommunications. As a result, the developed model can be used in cybersecurity audits or cybersecurity departments to the in-depth study of the information system and improve its resilience to hacker attacks.

The use of modeling to study each of the structural components of the warning system is determined to be appropriate. This is justified by the fact that it is impractical to conduct a real experiment, especially with the reproduction of cyber incidents, due to significant financial and labor costs. This approach is also effective when it is necessary to conduct an analysis of the designed system, which does not yet physically exist in this organization.

References

1. Kang B., Choo H. A deep-learning-based emergency alert system. *ICT Express*. 2016. Vol.2, Is. 2. Pp. 67–70. doi: 10.1016/j.icte.2016.05.001.
2. Hidayanti, Supangkat S. H. Designing a distribution emergency information service in earthquake post-disaster based on service computing system engineering. *In: Proceedings of the International Conference on ICT for Smart Society (ICISS).* Semarang, Indonesia. 2018. Pp. 1–6. doi: 10.1109/ICTSS.2018.8549969.
3. Berkunskyi Y., Knyrik K., Farionova T., Smykodub T. Using microservices in educational applications of IT-company. *In: Proceedings of the IEEE 1st Ukraine Conference on Electrical and Computer Engineering (UKRCON 2017).* Kyiv, Ukraine. 2017. Pp. 1208–1211. Article number 8100443. doi: 10.1109/UKRCON.2017.8100443.
4. Mishra S., Golias M. M., Thapa D. Work zone alert systems. *Technical Report RES2019-01*. Memphis, Tennessee : The University of Memphis,2021.
5. Kurtz J. A. Hacking wireless access points: Governmental context. *In book: Cracking, Tracking, and Signal Jacking*. Chapter 7. Elsevier, Burlington, MA : Syngress Publ. 2017. Pp. 93–107. doi: 10.1016/B978-0-12-805315-7.00007-3.
6. Garba F. A. The anatomy of a cyber attack: Dissecting the Cyber Kill Chain. *Scientific and Practical Cyber Security Journal (SPCSJ)*. 2019. Vol. 3, Is. 1. Pp. 29–44.
7. Tatam M., Shanmugam B., Azam S., Kannoorpatti K. A review of threat modelling approaches for APT-style attacks. *Heliyon.* 2021. Vol. 7, Is. 1. Article number e05969. doi: 10.1016/j.heliyon.2021.e05969.
8. Knight A. Threat modeling. *In book: Hacking Connected Cars: Tactics, Techniques, and Procedures*. Chapter 3. Hoboken, NJ : Wiley, 2020. doi: 10.1002/9781119491774.ch3.
9. Fruhlinger J. Threat modeling explained: A process for anticipating cyber attacks. Publ. 2020, April 15. URL: https://www.csoonline.com/article/3537370/threat-modeling-explained-a-process-for-anticipating-cyber-attacks.html.
10. Shostack A. 20 years of STRIDE: Looking back, looking forward. Publ. 2019, March 29. URL: https://www.darkreading.com/risk/20-years-of-stride-looking-back-looking-forward.
11. Zhang W. A distributed security situation evaluation model for global network. *CEUR Proceeding*. 2018. Vol. 2300. Pp. 245–248.
12. Burlachenko I., Zhuravska I., Davydenko Ye., Savinov V. Vulnerabilities analysis and defense based on MAS method in fast dynamic wireless networks. *In: Proceeding of the 4th IEEE International Symposium Wireless Systems within the IEEE International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IEEE IDAACS-SWS* 2018). Lviv, Ukraine. 2018. Pp. 98–102. doi: 10.1109/IDAACS-SWS.2018.8525692.
13. Merc L., Sobeslav V., Mikulecky P., Macinka M. Infrastructure Authentication, Authorization and Accounting solutions for an OpenStack platform. *In book: Mobile Web and Intelligent Information Systems. Lecture Notes in Computer Science.* Vol. 11673. London : Springer-Verlag, 2019. Pp. 123–135. doi: 10.1007/978-3-030-27192-3_10.
14. Types of wireless attacks. Publ. 2017, Jun 13. URL: https://blog.ct-networks.io/types-of-wireless-attacks-9b6ecc3317b9.
15. Prinetto P., Roascio G. Hardware security, vulnerabilities, and attacks: A comprehensive taxonomy. *CEUR Proceeding.* 2020. Vol. 2597. Pp. 12–23.

Sergii Bozhatkin **Сергій Божаткін**	Senior Lecturer of the Department of Computer Technologies and Information Security, Admiral Makarov National University of Shipbuilding, Mykolaiv, Ukraine, e-mail: sergii.bozhatkin@nuos.edu.ua https://orcid.org/0000-0002-4653-8880	старший викладач кафедри комп'ютерних технологій та інформаційної безпеки, Національний університет кораблебудування імені адмірала Макарова, Миколаїв, Україна
Viktoriia Guseva-Bozhatkina **Вікторія Гусєва-Божаткіна**	Senior Lecturer of Department of Automated Systems Software, Admiral Makarov National University of Shipbuilding, Mykolaiv, Ukraine, e-mail: GusevaBozh@meta.ua https://orcid.org/0000-0002-1117-3391	старший викладач кафедри програмного забезпечення автоматизованих систем, Національний університет кораблебудування імені адмірала Макарова, Миколаїв, Україна
Tetyana Farionova **Тетяна Фаріонова**	PhD on Engineering, Associate Professor, Director of the Educational and Scientific Institute of Computer Sciences and Project Management, Admiral Makarov National University of Shipbuilding, Mykolaiv, Ukraine, e-mail: tetyana.farionova@nuos.edu.ua https://orcid.org/0000-0003-3384-4712	канд. техн. наук, доцент, директор Навчально-наукового інституту комп'ютерних наук та управління проектами, Національний університет кораблебудування імені адмірала Макарова, Миколаїв, Україна.
Volodymyr Burenko **Володимир Буренко**	PhD Student of Department of Project Management, Admiral Makarov National University of Shipbuilding, Mykolaiv, Ukraine e-mail: volodymyr.burenko22@gmail.com https://orcid.org/0000-0002-0862-5879	аспірант кафедри управління проектами, Національний університет кораблебудування імені адмірала Макарова, Миколаїв, Україна
Bohdan Pasiuk **Богдан Пасюк**	PhD Student of Department of Automated Systems Software, Admiral Makarov National University of Shipbuilding, Mykolaiv, Ukraine e-mail: bwolverine44@gmail.com https://orcid.org/0000-0002-4634-4090	аспірант кафедри програмного забезпечення автоматизованих систем, Національний університет кораблебудування імені адмірала Макарова, Миколаїв, Україна

UDC 004.853:007.2
https://doi.org/10.31891/csit-2023-1-4

Ivan BURLACHENKO, Volodymer SAVINOV, Iryna ZHURAVSKA
Petro Mohyla Black Sea National University

THE ORGANIZING OF COMPETITIVE EVENTS USING MULTI-AGENT TECHNOLOGIES AND THE MODIFIED BORDA METHOD

The hackathons allow collecting at once on one site: the largest industrial companies of the country, technology vendors from the rapidly changing environment in the markets, young developers (including students), engineers with experience in the IT-sphere or specifically required technologies.

The current state of hackathon organizing stages has analyzed to improve the approach to increase the social inclusion of participants. Statistical metrics of vacancies occurrence probability during the period after the hackathon and employee turnover provided by hackathons' sponsors according to business domains were investigated.

The methods of determining the winner in different systems of competitive selection are considered. Particular attention is paid to the peculiarities of the tournament systems used in cybersport championships. The system of selection based on the modified Borda method, consisting of two or a maximum of three rounds and independent of the number of participants, is proposed.

In the paper, the Multi-Agent Sell Funnel Monitoring (MASFM) algorithm has described. MASFM algorithm allows searching sponsorship efficiently because it helps detect about 16–23% of new sponsors according to last 2 years statistics.

In the software architecture of the online hackathons' platform, a real scenario of increasing performance 15 times from 6 to 94 requests/sec was applied, which does not require serious refactoring and complex code changes. Besides, the steps mentioned above can reduce the cost of infrastructure like Heroku. The next functionality of the online hackathon platform will be possible thanks to the microservices architecture.

As the result, efficient software architecture has implemented and allow to decrease the maximum response time down to 3 seconds and the online hackathon platform's performance has increased from 71 to 94 requests per second.

Keywords: eSport event, organizing of hackathon, selection of teams, multi-agent system, algorithm for determining winners, grading procedure

Іван БУРЛАЧЕНКО, Володимир САВІНОВ, Ірина ЖУРАВСЬКА
Чорноморський національний університет імені Петра Могили

ОРГАНІЗАЦІЯ ЗМАГАНЬ З ВИКОРИСТАННЯМ МУЛЬТИАГЕНТНИХ ТЕХНОЛОГІЙ ТА МОДИФІКОВАНОГО МЕТОДУ БОРДА

Хакатони дозволяють зібрати на одному майданчику одночасно: найбільші промислові компанії країни, вендорів технологій зі стрімко мінливого середовища на ринках, молодих розробників (у тому числі студентів), інженерів з досвідом роботи в IT-сфері або у конкретно затребуваних технологіях.

Проаналізовано поточний стан етапів організації хакатону для вдосконалення підходу до підвищення соціальної інтеграції учасників. Досліджено статистичні показники ймовірності появи вакансій протягом періоду після хакатону та плинності кадрів, надані спонсорами хакатонів за сферами діяльності.

Розглянуто методи визначення переможця в різних системах конкурсного відбору. Особливу увагу приділено особливостям турнірних систем, які використовуються на чемпіонатах з кіберспорту. Детально розглянуто особливості застосування олімпійської системи організації IT-спортивних заходів, зважаючи на те, що кіберспорт неухильно наближається до того, щоб стати включеним у програму Олімпіади 2024. Пропонується система відбору переможців на основі модифікованого методу Борда, яка складається з двох або максимум трьох турів і не залежить від кількості учасників.

У статті описано алгоритм моніторингу воронки продажів із кількома агентами (MASFM). Алгоритм MASFM дозволяє ефективно шукати спонсорство, оскільки він допомагає виявити близько 16–23 % нових спонсорів за статистикою за останні 2 роки. У програмній архітектурі платформи онлайн-хакатонів реалізовано реальний сценарій збільшення продуктивності в 15 разів, що не потребує серйозного рефакторингу та складних змін коду. Крім того, застосовані послідовні кроки можуть знизити вартість такої загальновживаної в онлайн-хакатонах інфраструктури, як Heroku.

В результаті реалізована ефективна архітектура програмного забезпечення, що дозволяє зменшити максимальний час відповіді до 3 секунд, а продуктивність платформи онлайн-хакатону збільшити з 71 до 94 запитів на секунду. Подальше покращення функціональності платформи онлайн-хакатону можливо завдяки імплементації в розглянуту архітектуру мікросервісів.

Ключові слова: кіберспортивна подія, організація хакатону, відбір команд, мультиагентна система, алгоритм визначення переможців, порядок оцінювання

Introduction

Organizing an e-sports event is a complex process. Their influence on the various activities becomes possible due to the gamification of any processes, including in education. Gamification is the use of game practices and mechanisms in a non-game context during the learning process to engage users in problem-solving. The elements of the gamified process include joint actions to achieve their own goals, virtuality, countdown during the task for a limited time are elements of the gamified process, etc.

Hackathon is a way to find technological solutions for e-sports. The goal of the hackathon is to bring together students, developers, designers, data researchers, scientists, artists, 3D modelers, composers, managers with a variety of skills to develop joint projects and address specific challenges.

The hackathon allows you to collect four parties at once on one site: the largest industrial companies of the country, technology vendors from changing markets, young developers (including students), engineers with experience in the IT-sphere, or the specific required technologies. The point is that large companies come with their own tasks, and developers (in our case, students) try to show the concept of their solution at such hackathons. In a successful case, the participants in the hackathons receive contracts on the basis of which a company can be founded. Customers spend two or three days of their time answering questions, but they get a very good picture of technologies and many solution prototypes at once. It is for this reason that the hackathon, as an e-sports event, is an important stage in the educational process. The hackathon allows students to immerse themselves in the corporate culture of industrial companies for a certain period.

The Current State of Hackathon Organizing

There are different types of hackathons according to the type of participants (external in which anyone interested in the topic can take part and internal which are organized for a closed community of a particular company or organization), to the holding format (offline and online), etc. [1]. Offline, all participants gather in one place with round-the-clock access and spend the entire hackathon there. The organizers usually provide them with everything need (food, office, convenient places) so that the teams work on the task without being distracted.

The duration of the hackathon starts from 24 hours, the question arises how to sleep. And no way. At hackathons it is really customary to teach 100% and devote all your time to work on the project, so not everyone can sleep and not always.

During online hackathons, all processes from team building to pitching take place online and do not require the physical presence of participants. All interaction takes place either through special platforms for the hackathons or through separate online tools.

The advantages are that there can be many more participants in such online events than full-time participation, even from anywhere. It should be noted that the corona crisis has significantly changed the nature of the event market, displacing offline activity and significantly increasing the share of online activities, including in the world of hackathons [2]. The disadvantages are the difficulty of staying asleep when you sit at home in bed, much harder than when you work as a team on location.

The topics of hackathons are completely different. If at first, they were held only for IT specialists who gathered together for a group programming session, today hackathons take place in almost all professional fields and solve a variety of tasks. Hackathons conducted by Corporate Social Responsibility (CSR Ukraine) & UNFPA Ukraine (held by the United Nations) under the STEM Girls grant initiative are aimed at changing gender stereotypes in the IT industry [3]. The task of hackathons held under the program “ULEAD with Europe” financing by The European Union is to create information projects or specific programs with the help of information technology, which will become a starting platform for solving community problems [4].

The Entire Process of Organizing a Hackathon

Search for Participants and Advertising Campaigns Based on Multi-Agent Technologies

To find the initial set of sponsors the advertisement campaign needs to be started. Messenger channels and social networks are most suitable for audience coverage involved in the hackathon's workflow. For evaluating the effectiveness of the promoted project, you can use matrix algebra operations [5], fuzzy logic approach [6], evidence theory [7], or multi-agent technologies [8, 9]. During the participation, the metrics of every sponsor will be analyzed. Hackathon’s statistical metrics defined by equation (1).

$$F^{MASFM} = \begin{cases} \frac{V(P(T))}{WF_C(BD)} > FR_{avg}; \\ \frac{H^M(S_T,V_S)}{P(S_T)} > A_C\,. \end{cases} \quad (1)$$

In formula (1) F^{MASFM} is sponsor’s fitness function of the agent in Multi-Agent Sell Funnel Monitoring (MASFM). Sponsors are filtered according to the probability of vacancies V(P(T)) during the period after the hackathon. Employee turnover $WF_C(B)$ can be provided by sponsors according to business domains BD of IT company to define the sponsors’ vacancy fill rate that should be greater than the average value FR_{avg}. One of the key features investigated in F^{MASFM} is hackathon matching function $H^M(S_T, V_S)$ for the technology skills S_T and vacancies skills V_S. Participant’s skills $P(S_T)$ are important also and define the relations that should be greater than the average value A_C between all potential sponsors. The F^{MASFM} values have logged to the online hackathons platform database to be analyzed by organizers. Engaging developers is key to a successful hackathon because they know how to build applications. Online hackathon’s platform should also involve experts from the business domains, people from the communities, students, the wider the audience, the more creative solutions can be. The most expensive hackathon elements, they provide venue rent, food, and prizes. The algorithm for finding sponsors to support the hackathon and participants is presented in Algorithm 1. Also, engaging sponsors is influenced by what organizers can offer in exchange for support, so it is important to use MASFM to organize the hackathon.

ALGORITHM 1: Multi-Agent Sell Funnel Monitoring Algorithm

```
[participantsAgents, sponsorsAgents] = initializeMASFM(eventDate)
setupSponsorsFunnel(sponsorsAgents)
setupParticipantsFunnel(participantsAgents)
currentDate is inside monitoringPeriod
while currentDate is inside monitoringPeriod, do
        sponsorsAgents scan [probabilityOfVacancies, employeeTurnover] in monitoringPeriod
        for technologySkill in vacancySkills determined by sponsorsAgents, do
                for each participantSkill in list determined by participantsAgents do
                    calculate hackathon_matching_function_value
                    calculate sponsors_vacancy_fill_rate
                MASFMMetrics = aggregate(hackathon_matching_function_value ,
sponsors_vacancy_fill_rate)
          end
         analyze(MASFMMetrics) and sell sponsorship deal
end
```

Roles of Participants and Categories of Hackathons

To define the categories of hackathons in more detail, it is necessary to describe the roles of the participants. It will help increase the level of organization of the event process. Consider the main definitions used during the organization and conduct of the hackathon.

"The participant" is a specialist who has expressed a desire to participate in the Hackathon and received confirmation of participation from the organizers. "The team" consists of specialists who work together to create a project in Hackathon. "The team leader" is one of the team members who perform certain leadership functions: provides communication with mentors and organizers, represents his team at checkpoints and pitching.

"Themes" and "tracks" of the hackathon are specific areas within which projects will be developed. The tracks will bring together teams working on projects on the same topic. One track will correspond to one hackathon theme.

"Challenge" is a specific task presented by the organizers of the hackathon, in which you can also show your creativity. Teams can develop their unique project within one of the hackathon themes or work on a challenge.

"An expert mentor" is a mentor who has unique expertise, provides informational support and advice to hackathon teams in a particular field of knowledge. "A team mentor" is a mentor who moderates the work of teams. The team mentor will work with 3–5 defined teams.

"The judge" is an expert who will evaluate the projects developed by the participants. "The organizer" provides participants with everything necessary for productive work on the hackathon, monitors compliance with all rules and regulations (see Table 1).

The structure may differ slightly from hackathon to hackathon. The main and longest part of the hackathon is the teams' presentation of their ideas to the judges, who make decisions and award the winners.

Table 1

Detailed structure of the hackathon categories

Hackathon Type	Duration	Participants Roles	Sponsors Support	Judging
Long term	1–3 months	Judge, Participant, Mentor, Expert, Reviewer	Medium	Borda
Short-term	3–4 days	Judge, Participant, Mentor, Expert	Medium	Experts considered decisions
One-day	9–24 hours	Judge, Participant, Reviewer	Low	Experts considered decisions

The Software Architecture for Online Hackathons

The server side of the online hackathon platform is written by Spring Boot. Spring Boot is a project at the IO Execution level of the IO Spring Framework. With Spring Boot, the web application configurations are minimized as much as possible. Spring Boot supports embedded containers that allow web applications to run independently and without the need for a web server (see Fig. 1 below).

The web application of the hackathon platform shown in Figure 1 is working through the Java war command to export a war file to run on the Web Server. We used the "CLI Tools" to run spring scripts. Benefits of Spring Boot for the Hackathon Platform:

1) easily used to develop a Spring-based application with Java or Groovy Spring;
2) minimizes development time and raises productivity;
3) avoids writing a lot of boilerplate, Annotations, and XML configuration;
4) easily allows you to interact with Spring Boot applications with Spring Ecosystems like Spring JDBC, Spring ORM, Spring Data, Spring Security, etc.;

5) follows the "Default Configuration Principles" approach to minimize the time and effort invested in developing applications;

6) provides Embedded HTTP servers like Tomcat, Jetty ... to quickly and easily develop and test web applications;

7) provides CLI (Command Line Interface) tools for developing and testing Spring Boot (Java or Groovy) applications from the command prompt very easily and quickly;

8) provides many plugins for quickly developing and testing Spring Boot applications using Build tools like Maven and Gradle;

9) offers many plugins for easy handling of embedded databases and in-memory Databases.

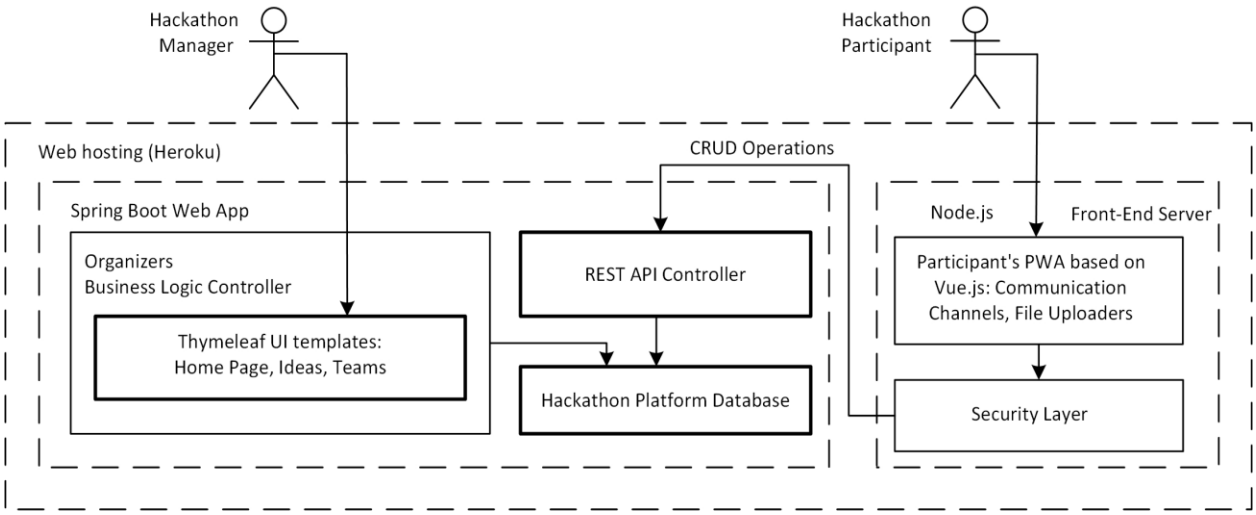


Fig. 1. Block diagram of the architectural solution for the hackathon online platform software

Heroku runs Spring Boot web applications inside one or more isolated "Dynos", which are virtual Unix containers that provide the necessary environment for your application. The dynos data is completely isolated and has an ephemeral file system (a "short-lived" file system is completely cleaned up and refreshed every time the dyno is restarted). Heroku internally uses a load balancer to distribute web traffic among all "web" dynos. Since the dynos are isolated, Heroku can scale the application horizontally by simply adding more dynos. The file system is ephemeral, so you cannot directly install the services your application needs (i.e., databases, queues, caching, storage, email services, etc.). Instead, Heroku provides services as independent "add-ons" either from Heroku and third parties. At the moment your application launches, dynos access services using the information contained in the configuration variables of your application.

To run your Heroku application, you need to be able to install the appropriate environment and dependencies. Developers interact with Heroku using a custom client application/terminal, which is very similar to a Unix bash script. It allows you to download code from a git repository, control running processes, view logs, and set configuration variables.

To get the hackathon platform application to work with Heroku, we needed to host our web application in the git repository, add the files listed above, connect the database add-on, and configure the settings to work correctly with static files. The Vue.js framework will allow you to create a responsive interface for both web and mobile platforms (Fig. 2).

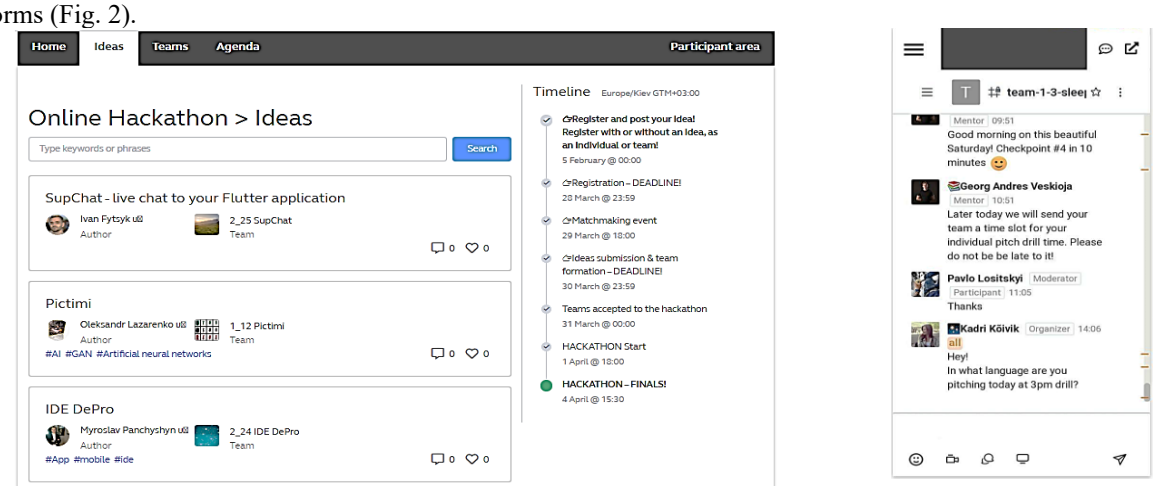
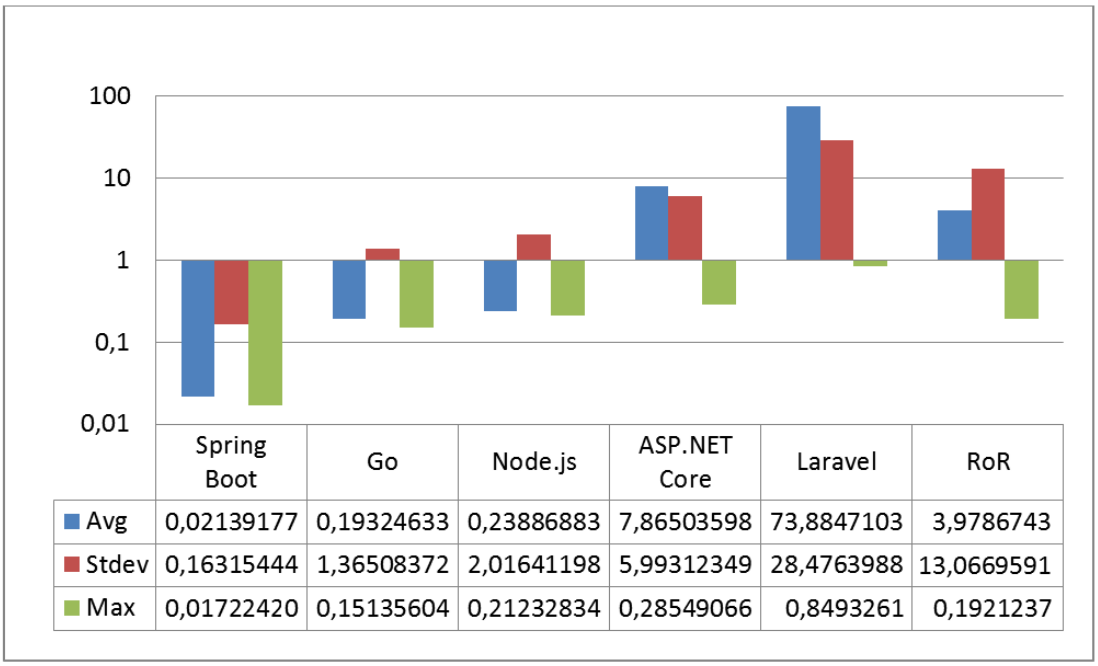


Fig. 2. The user interface of the online platform for hackathons

After completing all the necessary stages of the site of the hackathon platform, we can create a Heroku account, access the Heroku client, and use it to deploy a web platform for conducting hackathons.

The research used the JMeter application, which is designed to test the performance of web servers and is used as an automated tool for testing with test data, as well as a tool for functional testing of web applications, file servers, web servers, and even databases. During the experiments, important characteristics of JMeter were investigated. The JMeter application can be configured to simulate the N number of users or streams that load a specific web server or web application. JMeter measures web server performance by creating a simulated load on a web application. Moreover, several repeatable cycles can be done to get the average result, as well as see the test results in graphical and statistical form.

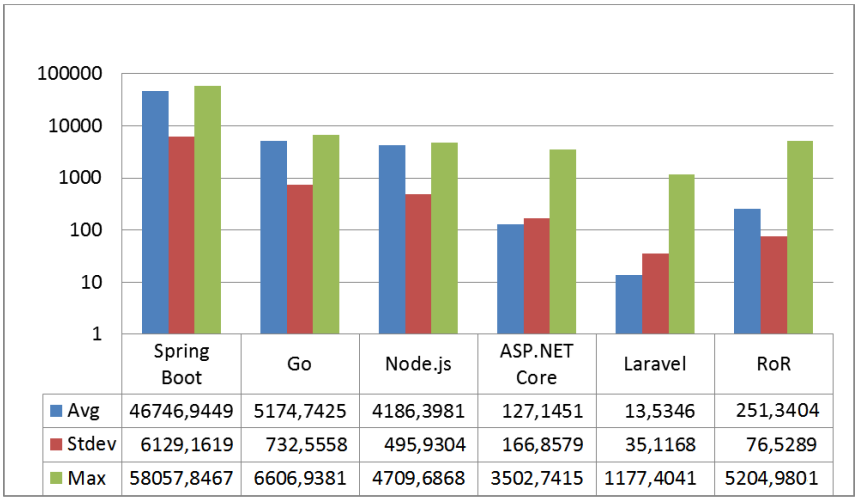
Based on the results, we can conclude that it is better to use clean technologies for small web applications. Note that ASP.NET Core used a large number of built-in libraries and MVC pattern. For clarity, a base-10 log scale was chosen to analyze the results was shown in Fig. 3.



	Spring Boot	Go	Node.js	ASP.NET Core	Laravel	RoR
Avg	0,02139177	0,19324633	0,23886883	7,86503598	73,8847103	3,9786743
Stdev	0,16315444	1,36508372	2,01641198	5,99312349	28,4763988	13,0669591
Max	0,01722420	0,15135604	0,21232834	0,28549066	0,8493261	0,1921237

Fig. 3. Web application request processing speed, ms

Web frameworks have specific conditions of use. If you have chosen a small web framework and need to develop a web application that is different from simple applications or the REST API, then you are likely to have problems with enhanced functionality, and vice versa - the redundancy of a full-featured, large web frameworks will cause financial the cost of placing content under high load shown in Fig. 4.



	Spring Boot	Go	Node.js	ASP.NET Core	Laravel	RoR
Avg	46746,9449	5174,7425	4186,3981	127,1451	13,5346	251,3404
Stdev	6129,1619	732,5558	495,9304	166,8579	35,1168	76,5289
Max	58057,8467	6606,9381	4709,6868	3502,7415	1177,4041	5204,9801

Fig. 4. The requests number to the web application per second

As a result, it was concluded that in conditions of hosting hardware resources it is advisable to develop web servers for hackathons based on the Spring Boot web framework.

The Methods of Determination and Awarding of Hackathon Winners

Existing algorithms for determining winners in sports

Over the years certain systems of selecting winners in different activities have developed. Sports are the oldest form of entertainment and competition. Since ancient Egypt and to this day people like to compete and win. Losing in any game, we strive to win the next one. eSports is already included in the Register of recognized sports in many countries. In September 2020 Ukraine also became one of more than 25 countries where cybersport is recognized as an official sport [10].

Therefore, it makes sense to consider the existing systems for determining winners in various sports (including Olympics) and analyze how they are suitable for judging in e-sports. Among the most popular systems of competitive selection of winners are the Olympic system ("playoffs") and the so-called "double-elimination system".

E-sports is steadily approaching to become included in the program of the 2024 Olympics [11], therefore it is advisable to consider in detail the features. Playoffs ensure that a winner is determined in a minimum number of rounds and promote a hard-fought tournament. Among the playoffs' advantages are the minimum number of games compared to other tournament variants and their "uncompromising" nature: there is no possibility and no point in a tie-break.

However, the playoffs are completely unsuitable for tournaments where it is important to ensure a fair distribution of all places, not just 1st – 3rd places.

First, in the playoffs, the distribution of places other than first is extremely influenced by the order in which the pairings are chosen. In a draw, the last places are allocated almost randomly: a weak competitor, who is drawn against an even stronger opponent, may easily rise above a stronger competitor in the first round [12].

Also, in a pure playoff, places other than 1st and 2nd can't be assigned at all. If it is necessary to specify the places occupied by participants, additional games have to be played, which is the greatest disadvantage of this selection system, because the main advantage of playoffs – speed – is lost.

In addition, a major disadvantage of playoffs is the rigid requirements on the number of participants. If that number does not comply with a norm, the only solution is to draw a draw to award technical wins or technical losses to some participants in the first round, which further increases the random factor's influence on the outcome of the tournament. The only alternative is to preempt a playoff tournament with a series of preliminary games for entry into the main tournament.

Thus, the Olympic system is built largely on a series of randomness, and the outcome of the competition is largely decided by lot, which is not fair to the other participants. Unlike familiar to us volleyball, soccer, or chess, in cybersport, everything is not so straightforward, and it becomes much more difficult to choose a winner.

So, in cybersport, a tournament system with elimination after two defeats, or as it is called, "two-consolation" or "double-elimination system" is commonly used [13]. This system is in contrast significantly to the simple Olympic system, in which a single defeat results in elimination. The double-elimination system is used in sports where it is easy to play twice as many matches as in the Olympic system, either due to the short match length or the large number of arenas running in parallel (automobile sports, darts, judo, etc.).

It is currently used in cybersports tournaments, including Dota 2 and CS: GO. In the tournament to two defeats played 2n–1 or 2n–2 games, depending on the outcome of the superfinal. This is at least twice as much as in the Olympic system, and the number of rounds at least one more. You can notice that more games must be played to reach the superfinals in the lower net than in the upper net.

Without detracting from the merits of this system, however, among its disadvantages should be noted the following:

1) special requirements for the number of participants (ideally a degree of twos). With the use of computer systems for competitions, there are usually no problems with the number of participants;
2) two sportsmen may face each other twice (and sometimes even three times);
3) most matches are played between outsiders and mediocre players and are of interest only to a narrow group of fans;
4) it is difficult to transport participants from one arena to another. In the Olympic system, for example, four stadiums host different branches of the tournament up to the quarterfinals, and then everyone is taken to one stadium where the semifinals and finals are played.

Thus, this system is also not the best way to select a winner and goes along with the Olympic system.

Based on the results of the analysis, it can be argued that traditional selection systems, even the double-elimination system, which is currently used in cybersport, are not universal. Against the backdrop of the sport's growing popularity, they carry with them a certain amount of inconvenience and unfairness. Therefore, it is advisable to develop a new algorithm for determining the winners of hackathons as a component of eSport.

Existing algorithms for determining winners in sports

Within this study, there is a development of the algorithm for a selection system that consists of two, at most three rounds and is completely independent of the number of participants.

The main goal of such system is to ensure the fairness of the competitive selection concerning all participants. The players will not be able to negotiate or bribe the judge, since the winner is selected by general voting.

The system is based on the Borda method of ranking. The Borda method is a voting system invented in 1770 by Jean-Charles de Borda to make the preferences of the electors more accurate when there are many candidates [14]. According to this method, the results of voting are expressed as the number of points scored by each of the candidates. Often this method does not give intuitively expected results when counting, thus preserving intrigue until the winners are announced [15].

This system can be compared to the selection system in parliamentary or presidential elections [16]. In this case, during the first round, each of the participants is assigned points in descending order from more to less. The system automatically counts the sum of points for each of the participants separately and arranges them in order from more to less respectively. Then a certain percentage of participants from the bottom (those with the lowest number of points) is "discarded". This percentage is not due to the Borda method, so the judge or the game administrator writes it into the program.

As a consequence of this selection, the participants who scored the highest number of points in total, go to the second round. Then the cycle repeats. The algorithm of the method is shown in Fig. 5.

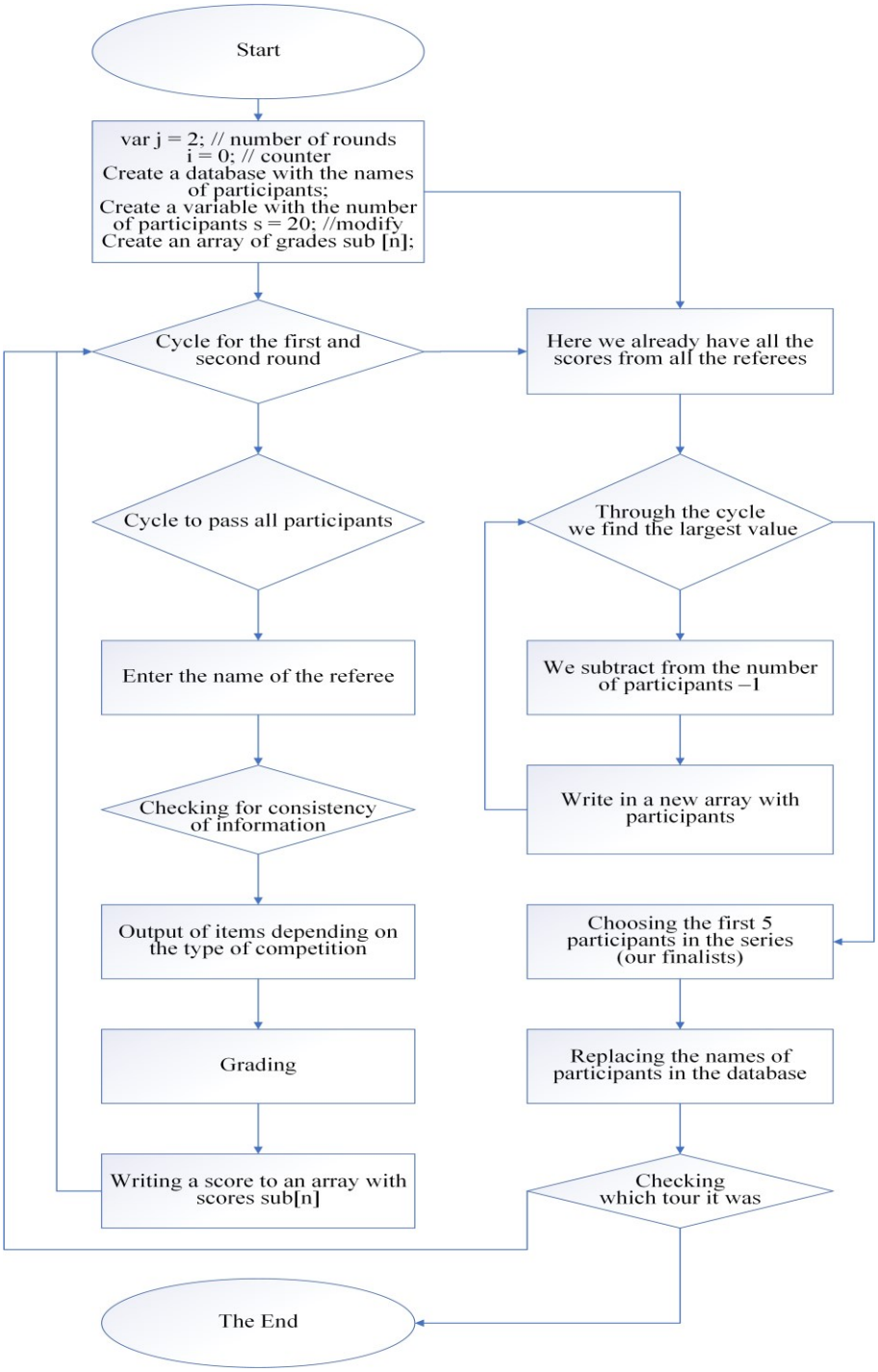


Fig. 5. The algorithm of the modified Borda method

The main advantages of this system for selecting a winner are:

1) the minimum number of rounds, and therefore the minimum amount of time to select a winner;
2) the impossibility of rigged matches, and thus fairness to all competitors;
3) intrigue until the end of the competition, up to the moment of the announcement of winners;
4) the possibility of fair assignment of prizes in any quantity.

The proposed system will allow a fair selection of the winner of any competition by general voting. And since cybersport is mostly a choice of a winner by voting by several judges, this system will help to calculate the number of points for each participant regardless of the others.

Conclusion

The proposed Multi-Agent Sell Funnel Monitoring (MASFM) algorithm allows to search sponsorship efficiently because according to last 2 years statistics about 16–23% of new sponsors was detected by MASFM. Now, when using Spring Boot, the maximum time is less than 3 seconds and the number of requests has increased from 71 to 94 requests per second. As expected, the percentage of errors increased to 29%. These all errors were by the hosting provider fault and test dyno container.

In the software architecture of the online hackathons' platform, a real scenario of increasing performance 15 times from 6 to 94 requests/sec was applied, which does not require serious refactoring and complex code changes. Besides, the steps mentioned above can reduce the cost of infrastructure like Heroku. The next functionality of the online hackathon platform will be possible thanks to the microservices architecture.

The applied system of grading based on modified Borda was assessed by the participants as fair enough. When analyzing the responses of the participants, negative responses due to the subjective conduct of judges decreased to 1–2%. The number of repeated participations in the hackathon is consistently above 70%.

Acknowledgments

The authors would like to thank the following grant sponsors of the annual event – hackathons held by the Faculty of Computer Science of Petro Mohyla Black Sea National University since 2016. There are IT companies whose offices are also located in Mykolaiv (Ukraine): FintechLab, IntroLab Systems, Room 4, Global Logic, Postindustria, MobiDev, Geeksforless, Mykolaiv IT Cluster (and give apologies to anyone we have missed acknowledging) [17].

References

1. Izvalov A., Nedilko S., Nedilko V. Comparison of game creation and engineering hackathons on the global and local levels. *In Proceedings of the Second International Conference on Game Jams, Hackathons, and Game Creation Events (ICGJ '17).* ACM International Conference Proceeding Series. 2017. Pp. 22–25. https://doi.org/10.1145/3055116.3055119.
2. Coronavirus: European Commission launches #EUvsVirus online hackathon. Publ. 2020, April 16. URL: https://news.liga.net/all/pr/koronavirus-evrokomissiya-zapustila-onlayn-hakaton-euvsvirus (Last accessed: 04.02.2023).
3. @divchataSTEM – Community. URL: https://www.facebook.com/divchataSTEM/ (Last accessed: 04.02.2023)
4. Open set for the 5th hackathon DEHACK 2019. URL: https://chmnu.edu.ua/vidkritij-nabir-na-p-yatij-hakaton-dehack-2019/ (Last accessed: 04.02.2023).
5. Trunov A., Fisun M., Malcheniuk A. The processing of hyperspectral images as matrix algebra operations. In Proceedings of the 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET). IEEE, Lviv-Slavske, Ukraine. 2018. Pp. 368–372. https://doi.org/10.1109/TCSET.2018.8336305.
6. Lavrynenko S., Kondratenko G., Sidenko I., Kondratenko Yu. Fuzzy logic approach for evaluating the effectiveness of investment projects. *In Proceedings of the 15th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT'2020).* IEEE, Zbarazh, Ukraine. 2020. No. 2. Pp. 297–300, Article number 9321880. https://doi.org/10.1109/CSIT49958.2020.9321880.
7. Shved A., Kovalenko I., Davydenko Ye. Method of detection the consistent subgroups of expert assessments in a group based on measures of dissimilarity in evidence theory. *Advances in Intelligent Systems and Computing. 1080 AISC.* 2020. Pp. 36–53. https://doi.org/10.1007/978-3-030-33695-0_4.
8. Tohoiev O., Burlachenko I., Zhuravska I., Savinov V. The monitoring system based on a multi-agent approach for moving objects positioning in wireless networks. *In Proceedings of the 3rd International Workshop on Computer Modeling and Intelligent Systems (CMIS-2020).* Zaporizhzhia, Ukraine, 2020. No. 2608. Pp. 79–90.
9. Zhuravska I., Obukhova K., Burlachenko I., Savinov V., Boiko A. Heatmaps for catering establishments web-applications available via MAS-improved wireless networks. *In Proceedings of the 5th IEEE International Symposium on Smart and Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS'2020).* IEEE, Dortmund, Germany. 2020. 6 p. Article number 9297107. https://doi.org/10.1109/IDAACS-SWS50031.2020.9297107.
10. Register of recognized sports in Ukraine: approved by the order of the Ministry of Youth and Sports of March 11, 2015, no. 639 (as amended by the order of Sept. 16, 2020, no. 1557). URL: https://zakon.rada.gov.ua/rada/show/v0639728-15#Text (Last accessed: 04.02.2023).
11. Cybersport at the Olympiad'2024? Publ. 21.02.2019. URL: https://futurenow.com.ua/kibersport-mozhe-zyavytysya-na-olimpiad/ (Last accessed: 04.02.2023).
12. Csato L. Two issues of the UEFA Euro 2020 qualifying play-offs. *International Journal of Sport Policy.* 2020. Vol. 12, Is. 1. Pp. 1–14. https://doi.org/10.1080/19406940.2020.1780295.
13. Aziz H., Gaspers S., Mackenzie S., Mattei N., Stursberg P., Walsh T. Fixing balanced knockout and double elimination tournaments. *Artificial Intelligence.* 2018. Vol. 262 (May 2018). https://doi.org/10.1016/j.artint.2018.05.002.
14. Felsenthal D. S., Nurmi H. Voting procedures under a restricted domain: An examination of the (In)vulnerability of 20 voting procedures to five main paradoxes. Springer, 2019. 92 p.
15. Regenwetter M., Grofman B. Approval voting, Borda winners and Condorcet winners: Evidence from seven elections. *Management Science.* 1998. Vol. 44, Is. 4. Pp. 520–533.
16. Electoral Systems. ACE electoral knowledge network. URL: https://aceproject.org/ace-ru/topics/es/ese/ese01/prezidentskie-vybory-dvuhturovaya-sistema (Last accessed: 04.02.2023).

17. The Faculty of Computer Science held the "Hackathon-2021", dedicated to the 25th anniversary of Petro Mohyla Black Sea National University. URL: https://chmnu.edu.ua/fakultet-komp-yuternih-nauk-proviv-hackathon-2021-prisvyachenij-25-richchyu-chnu-imeni-petra-mogili/ (Last accessed: 04.02.2023).

Ivan Burlachenko **Іван Бурлаченко**	Senior Lecturer of the Department of Computer Engineering, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine, e-mail: ivan.burlachenko2010@gmail.com https://orcid.org/0000-0001-5088-6709	старший викладач кафедри комп'ютерної інженерії, Чорноморський національний університет імені Петра Могили, Миколаїв, Україна
Volodymyr Savinov **Володимир Савінов**	PhD on Engineering, Associate Professor, Associate Professor of the Department of Computer Engineering, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine e-mail: volodymyr.savinov@chmnu.edu.ua https://orcid.org/0000-0002-0862-5879	канд. техн. наук, доцент, доцент кафедри комп'ютерних інженерії, Чорноморський національний університет імені Петра Могили, Миколаїв, Україна
Iryna Zhuravska **Журавська Ірина**	DrS on Engineering, Professor, Head of the Department of Computer Engineering, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine e-mail: iryna.zhuravska@chmnu.edu.ua https://orcid.org/0000-0002-8102-9854	доктор техн. наук, проф., зав. кафедри комп'ютерних інженерії, Чорноморський національний університет імені Петра Могили, Миколаїв, Україна

UDC 004.9: 004.05
https://doi.org/10.31891/csit-2023-1-5

Iryna ZASORNOVA, Tetiana HOVORUSHCHENKO, Oleg VOICHUR
Khmelnytskyi National University

STUDY OF SOFTWARE TESTING TOOLS ACCORDING TO THE TESTING LEVELS

Recently, software has been intensively used in almost all areas of business. Testing is an integral process of the software life cycle, during which it is proved that the software meets the specified requirements and needs of the customer, thereby ensuring the quality of the software. The article analyses the tools for software testing with their generalisation by levels of testing.

The study has shown that there are a number of studies aimed at reviewing and classifying software testing tools. The correct choice of software testing tools is one of the vital elements to ensure the quality of the entire project. However, most studies in the field of testing focus on describing testing methods without directly connecting to the tools that are based on these methods.

A specialist's approach to software testing requires additional information about the testing tools currently available. With the increasing complexity of software products and shorter development cycles, it is clear that manual testing cannot deliver the level of quality required by the market. Choosing the wrong testing tools for a project leads to inadequate quality measurements or tool changes during the project. Both wrong choice and change of testing tools during the development process affect the quality of the software product and, as a result, the success of the project as a whole.

The classifiers discussed in this paper can be used to select software testing tools appropriately. On the one hand, it can be useful for navigating a wide range of testing subjects, reducing the time required for specialists to find the right solution. On the other hand, it can be used as a short introduction to the rapidly developing field of testing and available testing tools for those who are not experts in this field. The classification can be applied to testing various software projects, depending on the type of software and development methodology.

Keywords: software, software testing, manual software testing, automated software testing, levels of software testing.

Ірина ЗАСОРНОВА, Тетяна ГОВОРУЩЕНКО, Олег ВОЙЧУР
Хмельницький національний університет

АНАЛІЗ ІНСТРУМЕНТІВ ТЕСТУВАННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ВІДПОВІДНО ДО РІВНІВ ТЕСТУВАННЯ

Останнім часом програмне забезпечення (ПЗ) інтенсивно використовується майже в усіх галузях підприємництва. Тестування є невід'ємним процесом життєвого циклу програмного забезпечення, під час якого доводиться, власне, відповідність ПЗ заданим вимогам і потребам замовника, тим самим забезпечується якість ПЗ. В статті проведено аналіз інструментів для тестування ПЗ з узагальненням їх по рівнях тестування.

Дослідження показало, що існує ряд досліджень, спрямованих на огляд і класифікацію інструментів тестування ПЗ. Коректний вибір інструментів для тестування ПЗ є одним із життєво важливих елементів для забезпечення якості усього проєкту. Проте більшість робіт у галузі тестування зосереджені на описі методів тестування без прямого підключення до інструментів, які базуються на цих методах.

Підхід фахівця до тестування ПЗ вимагає додаткової інформації про доступні на даний момент інструменти тестування. Із зростаючою складністю програмних продуктів та коротшими циклами розробки стає очевидним, що ручне тестування не може забезпечити рівень якості, необхідний для ринку. Неправильний вибір інструментів тестування для проєкту призводить до неадекватних вимірювань якості або заміни інструментів під час проєкту. Як неправильний вибір, так і зміна інструментів тестування в процесі розробки впливають на якість програмного продукту і, як наслідок, на успіх проєкту в цілому. Класифікатори, які розглянуто у роботі, можна використовувати для відповідного вибору інструментів тестування ПЗ. З одного боку, це може бути корисним для орієнтації в широкому предметному полі тестування, скорочуючи час, необхідний спеціалістам для пошуку вірного рішення. З іншого боку, його можна використати як короткий вступ до галузі тестування, що швидко розвивається, і доступних інструментів тестування для тих, хто не є експертом у цій галузі. Проведена класифікація може бути застосована для тестування різноманітних програмних проєктів, залежно від виду ПЗ та методології розробки.

Ключові слова: програмне забезпечення (ПЗ), тестування ПЗ, ручне тестування ПЗ, автоматизоване тестування ПЗ, рівні тестування ПЗ.

Introduction

Recently, software has been intensively used in almost all branches of business [1]. Testing is an integral process of the software life cycle, during which the compliance of the software with the specified requirements and needs of the customer is proven, thereby ensuring the quality of the software [2, 3].

Software testing can be manual or automated. In manual testing, testers perform tests manually without using any means of automation. Manual testing is a low-level and simple type of testing that does not require a lot of additional knowledge. However, before you can automate the testing of any application, you must first run a series of manual tests. Manual testing requires more effort, but without it, we cannot be sure whether automation is possible at all. One of the fundamental principles of testing is that 100% automation is impossible. Therefore, manual testing is a necessity [4].

Automated software testing is part of the testing process at the quality control stage in the software development process; it is a type of testing in which testing is performed using various automation tools and scripts. It uses software tools to execute tests and verify execution results, which helps reduce testing time and simplify the testing process.

Automated testing involves the use of special software to control the execution of tests and compare the expected and actual results of the program. This type of testing helps to automate activities that are often repeated, but which, at the same time, are necessary for maximum test coverage of the task. To compile automated tests, a QA specialist must be able to program. Automated tests are full-fledged programs that are simply designed for testing [4].

There are several main types of automated testing: Code-driven testing automation – testing at the level of software modules, classes and libraries (actually, automatic unit tests); graphical user interface testing automation (Graphical user interface testing) – a special program (testing automation framework) allows you to generate user actions – button presses, mouse clicks, and monitor the program's reaction to these actions – whether it meets the specification; automation of API (Application Programming Interface) testing – the software interface of the program, which is used to test interfaces intended for interaction, for example, with other programs or with the user. Here, again, as a rule, special frameworks are used [4].

The advantages of automated testing include:

– speed of execution of test cases is greater than with manual testing;

– lack of influence of the human factor in the process of execution of test cases;

– cost minimization during repeated execution of test cases (i.e. minimal human participation);

– the ability to perform such test cases that cannot be performed manually;

– the ability to collect, store, analyze, aggregate and present large volumes of data in a form convenient for human perception;

– the ability to perform low-level actions with the application, operating system, transmission channels, etc.

Disadvantages of automated tests include:

– the need for highly qualified personnel;

– high costs for complex automation tools, development and maintenance of test case code;

– automation requires more careful planning and risk management;

– complex selection of automation tools, the need to train staff (or find specialists);

– existing test cases may be unusable and outdated in the case of significant changes in requirements, changes in the technological domain, redesign of interfaces (both user and software), etc.

Currently, much attention is paid to the transition to fully automated software testing, as it provides better results, especially for large and super-large software projects, saves time and budget, allows to reduce the routine manual actions of the tester, and reduces the complexity of software testing [5]. Therefore, there is a rapid development of testing tools and methods. Both manual and automated testing can be used at different levels of testing, as well as be part of other types and types of testing.

Therefore, increasing the use of automated software testing is *an actual task* today.

The purpose of this study is to analyze software testing tools according to the level of testing.

Study of Software Testing Tools According to the Testing Levels

Software testing is usually performed by a team of testers or developers in accordance with set requirements (specifications). Depending on the type of software product, the testing method and appropriate software are chosen.

By level, testing is divided into Unit (first level), Integration (second level), and System (third level) [6]. However, acceptance testing (fourth level) is often added to this list [7]. In order to carry out quality testing, it is necessary to make a reasonable choice of testing tools. Using the classification, it is possible to choose the necessary testing tools.

Software testing tools are mostly designed to support one or more methods of testing a software product and can perform different tasks depending on the level of testing.

The conducted research showed that there is a sufficient number of testing tools designed for unit, integration, system, and acceptance testing.

Unit testing can be considered a basic level of testing with the ability to test several modules at the same time, which makes it possible to automate them. The free and open-source software xUnit is for the first level of testing like *Unit testing*, and depends on the programming language. For example, JUnit is designed for the Java programming language, NUnit for .NET, CppUnit, CUnit for C, C++, etc. [8]. In most cases, unit testing is performed by software product developers using automated tests.

The smallest number of tools was found for *Integration testing*. Testing tools for Unit testing (xUnit) are also used at the next (second) level – during Integration testing. This is related to the method of conducting Integration testing. Also using Rational Rose and Cantata++.

It is known that the behaviour of a software product changes to one degree or another after adding to it a new module that was tested at the first level of testing. Therefore, as part of Integration testing, *Regression testing* is carried out in order to check the functionality of the assembled parts (build) of the software product. Such tests are recommended to be performed using automation tools (Selenium, SilkTest, Rational Functional Tester and QEngine, etc.) since their set is constantly increasing during the development of a software product and manual testing is impractical.

System testing includes a set of the following tests: Recovery, Security, Stress and Performance testing [9]. A large number of tools have been identified for System testing, as this level of testing includes several more testing methods.

Acceptance testing is carried out at the last level before deployment. At this stage, a conclusion is made about the acceptance or rejection of the software product. At the same time, developers can be involved only to correct code errors, if any are found. As the project grows, the number of Acceptance testing will also increase. Therefore, for Acceptance testing, it is recommended to use automated testing tools, for example, FitNesse. FitNesse allows the customer of the software product to enter specially formatted input. Input data is interpreted and tests are created automatically. These tests are then run by the system and the output is returned to the customer. The advantage of this approach is fast feedback. The customer can write tests in the form of HTML tables and add, if necessary, additional text. The tool then analyzes the tables, runs the tests, and provides the results in an HTML document. FitNesse supports Java, but versions for C++, C#, Python, Ruby, Delphi, etc. are currently available.

In addition, it is worth noting that acceptance testing is not sufficiently automated today, as there is a lack of tools for this testing method. Therefore, the use of tools for system testing and acceptance testing is quite limited.

At the same time, it should be taken into account that the software for conducting automated testing is constantly improved, supplemented and expanded. Regarding the future solution, the tools presented in the study can be applied at different levels of testing, providing a practical demonstration of their use. Another possible improvement is to conduct a comparison between testing tools that belong to a similar group to present the advantages and disadvantages of a particular tool [10–11].

Consider the following types of testing. Thus, API testing checks the correctness of interaction between system components, UI testing tests the correctness of the graphical interface, and network protocol level testing checks the correctness of data transmission between computers. All these types of testing are important to ensure the quality of the software and its correct operation.

UI testing, i.e. the user interface, can be performed using the following tools: Abbot Java GUI Test Framework, Business Process Testing (BPT), QF-Test, cPAMIE, Jemmy, Quick Test Professional, Rational Functional Tester, Rational Robot, Selenium, SilkTest, TestComplete, TestPartner, TOSCA, Oracle Functional Testing.

The following tools are used for *API testing*: APL Sanity Autotest, Cantata++, CppTest, CppUnit, CUnit, DTM Data Generator, FitNesse, JProbe, JUnit, NUnit, JVerify, TestNG, VectorCAST/C++.

To *test the network protocol level* – in this case, the tool simulates the client part of the system that interacts with the object through network protocols. Tools: Conkormiq Qtronic, DTM DB Stress, HttpUnit, JMeter, LoadRunner, MessageMagic, NeoLoed, Nikto, OpenSTA, OpenTTCN Tester, Oracle Load Testing, QALoad, Rational App Scan, Rational Performance Tester, SilkPerformer, soapUI, The Grinder, WAPT, Weblnspect.

The listed tools have certain *disadvantages*. For example, although *Selenium* software is one of the most popular tools for automated website testing, it has some drawbacks:

– resistance to change: if the website under test undergoes changes in its structure or code, test scripts written using Selenium may become unusable, requiring the rewriting of tests;

– complexity of configuration: to use Selenium, you need to have basic knowledge of programming and environment configuration;

– limitations in operating systems: Selenium can only work on operating systems that support the browsers it works with;

– slow test execution speed: Selenium can sometimes lag during test execution, which can lead to a longer testing process;

– the need to support scripts: in order for Selenium to be able to work with some elements of the website, such as AJAX, it is necessary to have scripts that can be executed in the browser;

– limitation of liability: Selenium is not able to guarantee responsibility for the completeness of testing, so its use should be supplemented by manual testing and other software quality control methods.

Overall, although Selenium has its flaws, it is a pretty effective tool for automated website testing and performance testing.

Let's consider the advantages and disadvantages of the *Business Process Testing (BPT*) tool. Because of the complexity of testing business processes and the many applications involved, using code-based test automation is problematic. Coded test automation takes time to develop and test. With BPT and multiple scenario testing, the time spent building automated coded testing makes it slow and a significant barrier for organizations. Testing multiple applications requires expertise and knowledge of each application. Coders who develop automation do not have deep applied knowledge. With BPT and multiple applications, this increases testing slowdowns. Advantages of BPT: eliminates the need to create a separate automation system; automated testing becomes structured using business components; reduces the effort required to write and maintain test automation scripts BPT is independent of the detailed test script; high reusability with data-driven components: testers do not need technical knowledge in automation. Disadvantages of BPT: an additional license for the BPT Framework must be purchased for test scripts; The BPT Framework can only be used if you have access to Application Lifecycle Management (ALM).

QF-Test from *Quality First Software* is a cross-platform graphical user interface test automation software tool specializing in Java/Swing, SWT, Eclipse and RCP plug-ins, Java applets, Java Web Start, ULC and cross-browser static and dynamic test automation web applications. A sophisticated recognition mechanism provides extreme serviceability and low maintenance costs, which is the most important factor in software testing automation.

The results of the analysis of tools for software testing with their generalization by testing levels are presented in Table 1.

Table 1

Comparative analysis of tools for automated software testing

№	Tools	Criteria					
		stable to change	**easy to set up**	**supports various OS**	**test performance speed**	**script support**	**URL**
				UI testing			
1	Abbot Java GUI Test Framework	stable	difficult	works with all platforms that support Java	low	no	http://abbot.sourceforge.net/
2	Business Process Testing (BPT)	unstable if business processes change regularly	complex in case when many business processes need to be configured	may be limited in the OS on which its components can be used	low	may require support for scripts used to configure tests and automate their execution	http://www.hp.com/
3	QF-Test	stable, but if the PP changes frequently, the tests may require frequent modification	difficult, but if the PP changes frequently, the tests may require frequent modification	supports, but some functions may be limited	low	yes	http://www.qfs.de/en/qftest/index.html
4	cPAMIE	stable if the page does not contain dynamic content	medium difficulty	Windows, Linux	very low	can support JavaScript	http://pamie.sourceforge.net/
5	Jemmy	very stable	medium difficulty, requires some additional components and libraries NetBeans RCP	Windows, Linux, Mac OS X	low	supports based on Java	https://jemmy.dev.java.net/
6	Quick Test Professional	stable, but with frequent requires checking and correction	difficult	Windows	low	scripting support is required	http://www.hp.com/
7	Rational Functional Tester	stable	easy, but you need to have experience in test automation	supports	high, but depends on the volume of test scenarios and the complexity of the application	uses Java scripts	http://www.ibm.com/
8	Rational Robot	stable in the	easy	Windows	high	scripting	http://www.ibm.com/

		absence of significant changes in the user interface				support is required	
9	Selenium	stable	easy	supports	high, but depends on the size of the web page, the number of elements	yes	http://seleniumhq.org/
10	SilkTest	unstable	easy	supports	high, but depends on the complexity of software and the number of tests	yes	http://www.borland.com/
11	TestComplete	very stable	easy	supports	high	yes	http://automatedqa.com/products/testcomplete/
12	TestPartner	stable	easy	Windows	high	yes	http://www.microfocus.com/
13	TOSCA	stable	easy	supports	high	yes	http://www.tricentis/com/
14	Oracle Functional Testing	very stable	easy	supports	high	yes	http://www.oracle.com/
15	Canoo WebTest	stable if the structure of the web pages or web application does not change dramatically	easy	Windows, Linux, Mac OS X	high, but depends on the size of the test suites and the complexity of the web application	yes	http://www.testingtoolsguide.net/tools/canoo-webtest/
16	QEngine	stable	easy	Windows, Linux, Mac OS X	high, but depends on the size of the test sets and the complexity of the software	yes	http://www.manageengine.com/products/qengine/index.html
				API testing			
17	APL Sanity Autotest	stable	easy	Windows, Linux, Mac OS X	high	yes	http://ispras.linux-foundaton.org/index.php/API_Sanity_Autotest
18	Cantata++	stable if additional functions do not require significant changes to the source code	easy	Windows, Linux, Mac OS X	high	yes	http://www.ipl.com/products/tools/pt400.uk.php
19	CppTest	stable	easy	supports	very high	no	http://cpptest.sourceforge.net/
20	CppUnit	stable	easy	supports	high	no	http://cppunit.sourceforge.net/
21	CUnit	stable	easy	supports	very high	no	http://cunit.sourceforge.net/
22	DTM Data Generator	stable	easy	Windows or with the help of virtual machine	high	no	http://www.sqledit.com/dg/

23	FitNesse	very stable	easy	Windows, Linux, Mac OS X	low	так	http://www.fitnesse.org/
24	JProbe	unstable	easy	supports	high	yes	http://www.quest.com/jprobe/
25	JUnit	stable	easy	Windows, Linux, Mac OS X	high	no	http://www.junit.org/
26	NUnit	stable	easy	Windows	very high	yes	http://www.nunit.org/
27	TestNG	very stable	easy	Windows, Linux, Mac OS X	high	yes	http://testing.org/
28	VectorCAST/ C++	stable	easy	Windows, Linux, Mac OS X	high	yes	http://www.vectorcast.com/software-testing-products/c++-unit-testing.php
29	JVerify	stable	easy	supports	high	no	https://jverify.us/
Network protocol level testing							
30	Conkormiq Qtronic	stable	easy	Windows, Linux, Mac OS X	high	yes	http://www.conformiq.com/qtronic.php
31	DTM DB Stress	stable	easy	Windows, Linux, Mac OS X	high	yes	http://www.sqledit.com/stress/index.html
32	HttpUnit	stable, if the server code changes, the tests may need to be updated	easy	Windows, Linux, Mac OS X	high	yes	http://httpunit.sourceforge.net/
33	JMeter	very stable	easy	Windows, Linux, Mac OS X	high	yes	http://jakarta.apache.org/jmeter/
34	LoadRunner	stable, but changing existing tests can be quite difficult	easy	Windows	high	yes	https://www.hp.com/
35	MessageMagic	stable, but depends on external libraries	easy	supports	high	yes	http://www.elvior.com/messagemagic/
36	NeoLoed	very stable	easy	supports	very high	yes	http://www.neotys.com/
37	Nikto	stable	easy	Windows, Linux, Mac OS X	very high	no	http://cirt.net/nikto2
38	OpenSTA	stable	difficult	Windows, Linux, Mac OS X	low	yes	http://www.opensta.org/
39	OpenTTCN Tester	stable	easy	Windows, Linux, Mac OS X	high	yes	http://www.opentten.com/
40	Oracle Load Testing	stable	easy	Windows, Linux	high	yes	http://www.oracle.com/
41	QALoad	stable except for new versions	easy	Windows, Unix	high, but depends on the complexity of the tests	yes	http://www.microfocus.com/

					and the amount of data		
42	Rational App Scan	stable if the web application does not undergo significant changes in structure or design	easy	Windows, Linux	high	yes	http://www.ibm.com/
43	Rational Performance Tester	stable	easy	Windows, Linux, Mac OS X	high	yes	http://www.ibm.com/
44	SilkPerformer	stable if changes are not very significant and the test script should not be completely reworked	easy	supports	high	yes	http://www.borland.com/
45	soapUI	stable	easy	Windows, Linux, Mac OS X	high	yes	http://www.soapui.org/
46	The Grinder	stable	difficult	Windows, Linux, Mac OS X	high for small projects	yes	http://grinder.sourceforge.net/
47	WAPT	stable if the changes are not very significant	easy	Windows, Linux, Mac OS X	high	yes	http://loadtestingtool.com/
48	WebInspect	stable if the changes are not very significant	easy	Windows, Linux, Mac OS X	high, but depends on the complexity of the web application, the number of tests and the amount of data	yes	http://www.hp.com/

Results & Discussion

When performing an analysis of testing tools, the following was found: for UI testing – 16 tools; for API testing – 13 tools; for testing the network protocol level – 19 tools.

In general, there are 9 tools for Unit testing, 4 for Integration testing, 15 for System testing, and 20 for Acceptance testing. The diagram of the distribution of tools depending on the level of testing is presented in Fig. 1.

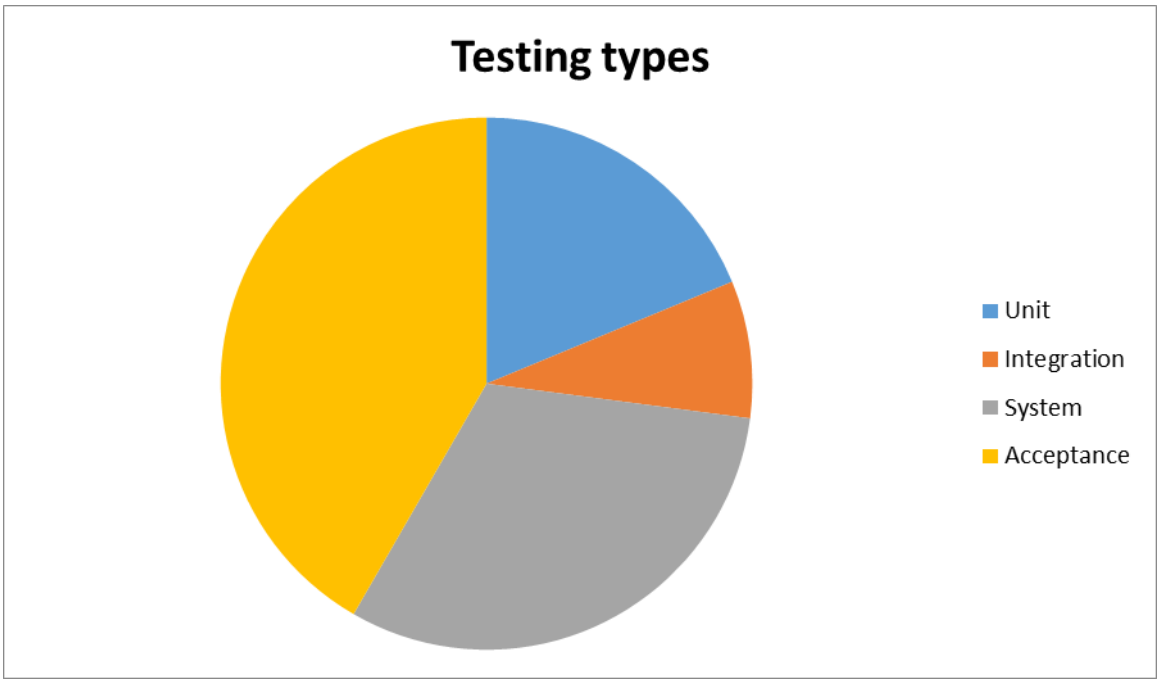


Fig. 1. Diagram of the distribution of tools depending on the level of testing

The conducted research can be used when choosing tools for testing software products.

Conclusions

Recently, software has been intensively used in almost all branches of business. Testing is an integral process of the software life cycle, during which the compliance of the software with the given requirements and needs of the customer is proven, thereby ensuring the quality of the software. The article analyzes software testing tools with their generalization by testing levels.

The study showed that there are a number of studies aimed at reviewing and classifying software testing tools. The correct choice of tools for software testing is one of the vital elements for ensuring the quality of the entire project. However, most of the work in the field of testing focuses on describing test methods without directly connecting to the tools that are based on these methods.

A specialist's approach to software testing requires additional information about currently available testing tools. With the increasing complexity of software products and shorter development cycles, it is becoming apparent that manual testing cannot provide the level of quality required by the market. Choosing the wrong test tools for a project leads to inadequate quality measurements or tool replacement during the project. Both the wrong choice and the change of testing tools during the development process affect the quality of the software product and, as a result, the success of the project as a whole.

The classifiers considered in the work can be used for the appropriate selection of software testing tools. On the one hand, it can be useful for orientation in a wide subject field of testing, reducing the time needed by specialists to find the right solution. On the other hand, it can be used as a brief introduction to the rapidly developing field of testing and available testing tools for those who are not experts in the field. The conducted classification can be used for testing various software projects, depending on the type of software and development methodology.

References

1. L. Li, Y. Tian. Application of Data Guidance Site Generation Technology in the Cloud Platform Supporting the Construction of Subject Teams in Finance and Economics Applied Universities. 2022 International Conference on Edge Computing and Applications: Proceedings (Tamilnadu (India), 2022). Tamilnadu, 2022. Pp. 19-22.
2. T. Hovorushchenko. Methodology of Evaluating the Sufficiency of Information for Software Quality Assessment According to ISO 25010. Journal of Information and Organizational Sciences. 2018. Vol. 42. No.1. Pp. 63-85.
3. T. Hovorushchenko, O. Pavlova, D. Medzatyi. Ontology-Based Intelligent Agent for Determination of Sufficiency of Metric Information in the Software Requirements. Advances in Intelligent Systems and Computing. 2020. Vol. 1020. Pp. 447-460.
4. Manual and automated testing. URL: https://qalight.ua/baza-znaniy/ruchne-ta-avtomatizovane-testuvannya/.
5. M. Rennhard, M. Kushnir, O. Favre, D. Esposito, V. Zahnd. Automating the Detection of Access Control Vulnerabilities in Web Applications. SN Computer Science. 2022. Vol. 3. Paper no. 376.
6. M. Surendra Naidu. Classification of defects in software using decision tree algorithm. International Journal of Engineering Science and Technology. 2013. Vol. 5. Paper no. 06.
7. M. Felderer, P. Zech, R. Breu, M. Büchler, A. Pretschner. Model-based security testing: a taxonomy and systematic classification. Software Testing, Verifiation & Reliability. 2016. Vol. 26. Issue 2. Pp. 119-148.
8. S. Nair, J. L. de la Vara, M. Sabetzadeh, L. Briand. Classification, Structuring, and Assessment of Evidence for Safety - A Systematic Literature Review. 2013 IEEE Sixth International Conference on Software Testing, Verification and Validation: Proccedings (Luxembourg (Luxembourg), 2013). Luxembourg, 2013. Pp. 94-103.

9. M. Felderer, B. Agreiter, P. Zech, R. Breu. A Classification for Model-Based Security Testing. The Third International Conference on Advances in System Testing and Validation Lifecycle: Proceedings (Madrid (Spain), 2011). Madrid, 2011. Pp. 109-114.

10. S. Kumar, S. Mane, S. Mali. A Comparative Study of Machine Learning Algorithms for Software Quality Classification. Journal of King Saud University - Computer and Information Sciences. 2021. Vol. 33. Issue 3. Pp. 308-315.

11. M. Gokhale, S. Shukla. Improving the Software Development Process with SWEBOK and Agile Methodologies. The 4th International Conference on Computing Methodologies and Communication: Proceedings (Erode (India), 2021). Erode, 2021. Pp. 476-484.

Iryna Zasornova **Ірина Засорнова**	Associate Professor of Computer Engineering & Information Systems Department, Khmenlnytskyi National University https://orcid.org/0000-0001-6655-5023 e-mail: izasornova@gmail.com	Кандидат технічних наук, доцент, доцент кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет
Tetiana Hovorushchenko **Тетяна Говорущенко**	DrSc (Engineering), Professor, Head of Computer Engineering & Information Systems Department, Khmenlnytskyi National University https://orcid.org/0000-0002-7942-1857 e-mail: govorushchenko@gmail.com	Доктор технічних наук, професор, завідувач кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет
Oleg Voichur **Олег Войчур**	Lecturer of Computer Engineering & Information Systems Department, Khmenlnytskyi National University https://orcid.org/0000-0001-8503-6464 e-mail: o.voichur@gmail.com	Асистент кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет

UDC 004.85
https://doi.org/10.31891/csit-2023-1-6

Oleksandr IERMOLAIEV, Inessa KULAKOVSKA
Petro Mohyla Black Sea National University

IMPROVING THE QUALITY OF SPAM DETECTION OF COMMENTS USING SENTIMENT ANALYSIS WITH MACHINE LEARNING

Nowadays, people spend more and more time on the Internet and visit various sites. Many of these sites have comments to help people make decisions. For example, many visitors of an online store check a product's reviews before buying, or video hosting users check at comments before watching a video. However, not all comments are equally useful. There are a lot of spam comments that do not carry any useful information. The number of spam comments increased especially strongly during a full-scale invasion, when the enemy with the help of bots tries to sow panic and spam the Internet. Very often such comments have different emotional tone than ordinary ones, so it makes sense to use tonality analysis to detect spam comments. The aim of the study is to improve the quality of spam search by doing sentiment analysis (determining the tonality) of comments using machine learning. As a result, an LSTM neural network and a dataset were selected. Three metrics for evaluating the quality of a neural network were described. The original dataset was analyzed and split into training, validation, and test datasets. The neural network was trained on the Google Colab platform using GPUs. As a result, the neural network was able to evaluate the tonality of the comment on a scale from 1 to 5, where the higher the score, the more emotionally positive the text and vice versa. After training, the neural network achieved an accuracy of 76.3% on the test dataset, and the RMSE (root mean squared error) was 0.6478, so the error is by less than one class. With using Naive Bayes classifier without tonality analysis, the accuracy reached 88.3%, while with the text tonality parameter, the accuracy increased to 93.1%. With using Random Forest algorithm without tonality analysis, the accuracy reached 90.8%, while with the text tonality parameter, the accuracy increased to 95.7%. As a result, adding the tonality parameter increased the accuracy for both models. The value of the increase in accuracy is 4.8% for the Naive Bayes classifier and 4.9% for the Random Forest.

Keywords: sentiment analysis, spam detection, neural network, text analyze, Python.

Олександр ЄРМОЛАЄВ, Інеса КУЛАКОВСЬКА
Чорноморський національний університет імені Петра Могили

ПОКРАЩЕННЯ ЯКОСТІ ПОШУКУ СПАМУ В КОМЕНТАРЯХ ЗА ДОМОГОГОЮ АНАЛІЗУ ТОНАЛЬНОСТІ З ВИКОРИСТАННЯМ МАШИННОГО НАВЧАННЯ

У наш час люди все більше і більше проводять часу в Інтернеті та відвідують різноманітні сайти. Багато з цих сайтів мають коментарі, що допомагають людям приймати рішення. Так, багато відвідувачів інтернет-магазину дивиться на відгуки до товару перед покупкою, а користувачі відеохостингів часто орієнтуються на коментарі перед переглядом. Проте не всі коментарі однаково корисні, досить часто можна зустріти спам-коментарі які не несуть жодної корисної інформації. Особливо сильно зросла кількість спам-коментарі під час повномасштабного вторгнення, коли ворог за допомогою ботів намагається посіяти паніку та заспамити Інтернет простір. Часто такі коментарі відрізняються за емоційним забарвленням від звичайних, тому існує сенс використовувати аналіз тональності для їх виявлення. Метою дослідження є покращення якості пошуку спаму за допомогою визначення тональності коментарів з використанням машинного навчання. В результаті було обрано LSTM нейромережу та датасет для її навчання та перевірки. Було описано три метрики для оцінки якості нейромережі, а датасет було проаналізовано та розбито та навчальну, валідаційну та тестову вибірки. Навчання нейромережі відбувалося на платформу Google Colab з використанням GPU. У результаті нейромережа змогла оцінювати тональність коментаря по шкалі від 1 до 5, де чим вище оцінка – тим більш емоційно-позитивний відгук і навпаки. Після навчання нейромережа досягла точності у 76.3% на тестовому датасеті, а серденя квадратична помилка становила 0.6478, що позначає що нейромережа помиляється менше ніж на один клас. При використанні алгоритму наївного байєсівського класифікатора без аналізу тональності, точність склала 88.3%, тоді як з параметром тональності тексту точність зросла до 93.1%. При використанні алгоритму випадкового лісу без аналізу тональності, точність склала 90.8%, тоді як з параметром тональності тексту точність зросла до 95.7%. В результаті що додавання параметру тональності підвищило точність для обох моделей. Значення приросту точності становить 4.8% для наївного байєсівського класифікатора та 4.9% для випадкового лісу.

Ключові слова: аналіз тональності, пошук спаму, нейромережі, аналіз тексту, Python.

Introduction

The main feature of modern AI algorithms is that they can "accumulate experience". In this way, they are able to solve some tasks with informal conditions, which is not able to do any productive, but rigidly programmed computer system.

A neural network consists of a system of connected and interacting simple processors called neurons. They are usually quite simple, especially compared to the processors used in personal computers. Each neuron of such a network connects only with signals that it receives and signals that it sends to other neurons [1]. Nevertheless, being connected in a rather large network, such individual simple parts together are capable of performing quite complex tasks.

The fields of application of neural networks are quite diverse – these are text and speech recognition, semantic search, expert systems and decision support systems, stock price prediction, security systems, text analysis, etc. This study considers an example of using a neural network to analyze the tone of comments.

Probably, there are tasks for neural network in each subject area. Here is a list of individual areas where the solution of this kind of tasks is of practical importance already now: economy and business, medicine, communication

and the Internet, production automation, political and sociological technologies, security and security systems, input and processing of information, geological exploration.

Classification tasks are understood as tasks for dividing a set of input signals into a predetermined number of classes [2]. After training, such a network is able to determine to which class the input signal belongs. In some varieties, the neural network can signal that the input signal does not belong to any of the selected classes – this is a sign of the appearance of new data that is not in the training sample, or of incorrect input data.

The field of text tonality analysis is quite new, and new technologies appear in it every year. Currently, one of the most popular tools for this is Cognitive Service from Microsoft Azure (Fig. 1). However, it has several disadvantages. One of them is that it is part of the Microsoft Azure cloud platform, so in order to use it, you need to familiarize yourself with the basics of this platform and configure access. In addition, the neural networks for this service were trained on the texts of articles from the Internet, so this service will have less accuracy on reviews. Instead, the field of spam detection is popular and there are many proven solutions on the market. One of them uses Google Gmail to filter spam in emails. However, it is not known whether it use tonality analysis for spam detection.

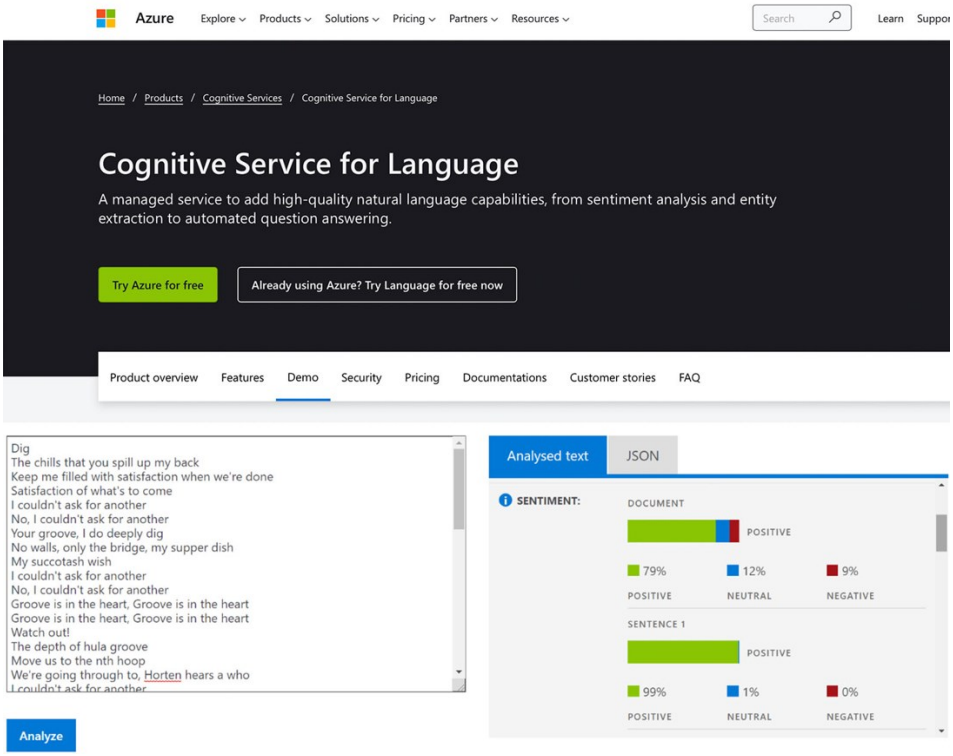


Fig.1. Similar system – Microsoft Azure Cognitive Service

One of the tasks that artificial intelligence can solve is the analysis of the tonality of the text (sentiment analysis) [3]. This means analyzing the emotional coloring of the text, taking into account its context. Tonality is the emotional attitude of the author of the statement towards some object expressed in the text. The emotional component expressed at the level of a lexeme or communicative fragment is called lexical tonality (or lexical sentiment). The tonality of the entire text as a whole can be determined as a function (in the simplest case, as a sum) of the lexical tonalities of its constituent units (sentences) and the rules for their combination. The main goal of tonality analysis is to find thoughts in the text and identify their properties. In this research, the tonality analysis was used to assign the text to one of five classes. So, it turns the task into classification problem.

Data preparation and neural network creation

We decided to use recurrent neural network, especially a structure called LSTM. This choice was justified by the ability of these neural networks to remember the last analyzed words [4], and therefore to remember the context.

Firstly, it was necessary to choose a dataset. The right dataset is a very important part of training a neural network. It must follow next conditions:

– have a sufficient amount of data;
– these data must be up-to-date;
– the dataset must be balanced, have approximately the same amount of data in each class;
– must be collected from a reliable source.

In addition, the dataset needs to be divided into several parts. Train sample is a sample for which training takes place. Each training iteration will process this dataset, or a certain part of it. This part of the dataset should be

the largest. It is very important that it meets the requirements described above, as this part of the dataset will be responsible for the accuracy of the model.

When model is built according to the training sample, then the quality assessment of this model, made on the same sample, turns out to be more optimistically [5, 6]. This phenomenon is called overtraining and in practice it occurs very often. A good empirical assessment of the quality of the constructed model is provided by its verification on independent data that were not used for training. Therefore, as described in it, in order to avoid the phenomenon of overtraining, it is necessary to check the neural network on another part of the data, which is called validation. This part of the dataset is significantly smaller than the training sample.

A test or control dataset is a sample that evaluates the quality of the constructed model. If the training and validation samples are fed to the model input multiple times, it is very important that the network sees the test dataset as few times as possible. It is on the basis of this sample that the final accuracy of the neural network will be evaluated.

To solve the task of comment analysis, we chose a fairly well-known dataset [7], which meets all the criteria described above. The dataset itself consists of 700,000 reviews that were taken from the famous USA site yelp.com. In addition to the text, each comment has a tonality rating from 1 to 5.

Preprocessing and training took place in Jupyter Notebook. It is an interactive computing environment that allows users to write, run, and share code, and display and manipulate data in an interactive and collaborative manner. Jupyter Notebook supports a variety of programming languages, including Python. One of the key strengths of Jupyter Notebook is its ability to interactively calculate each step. In addition, it has powerful built-in data visualization capabilities.

The input data has been cleaned and tokenized. In cleaning stage, we removed all non-alphanumeric characters from each part of the text. Tokenization is the process of breaking text into smaller pieces called tokens [8]. In this case, the tokens will be individual words. Therefore, further each comment is divided into separate words. We use an English dictionary from nltk library to do tokenization. It is used to bring all the words of the main form. This process is called stemming and is an important part of text normalization [9]. Since there is no ideal algorithm for finding the basic form, we use English dictionary from nltk that containing words and their various forms. This stemming method is called table lookup. The advantages of this method are the simplicity, speed and convenience of handling exceptions for each language. The disadvantages include the fact that the search table must contain all forms of words, which means that the algorithm will not work with new words.

After that, we convert each word into a vector using "word2vec" algorithm. It accepts a large text corpus as input data and assigns a vector to each word, giving the coordinates of the words at the output. For this task, the dictionary contained 2000 words that were most frequently encountered among all comments. All fewer common words were replaced with the key "UNK". During text analysis, large datasets can contain thousands or even millions of unique words, many of which are irrelevant or do not add significant information to the analysis. When building a model, it is important to keep the number of parameters to a minimum, as each new parameter complicates the model and increases the risk of overtraining. Conversely, by limiting the number of parameters, it is easier and faster to train the model. It can also reduce the computational cost of training the model, as well as reduce the memory required to store the feature matrix. This is especially important when training a neural network on a GPU in Google Colab, as there are limits on the maximum time of use of computing resources.

We created a network accepts a vector of a comment text and has five outputs. Each one matches rating from 1 to 5. The output with the highest score is the most likely class number. After several training iterations, a dropout method was added. This is a regularization method used in neural networks to prevent overtraining. It works by randomly removing a certain percentage of neural network connections during training. This forces the network to create redundant correct connections, making it more robust and less likely to overtrain. The dropout value is a hyperparameter and is determined by experiment [3]. A value of 0.2 was chosen for this neural network.

The entire learning process was divided into epochs. In each epoch, the training dataset is divided into groups of 2000 comments. Since the training sample contains 80% of 650,000 comments, then the total count of groups is $\frac{0.8 \cdot 650000}{2000} = 260$. Validation occurs at the end of each epoch. There are three metrics used in validation: the logistic loss function, the precision and the root mean square error (RMSE). Training and verification of the neural network took place on the Google Colab cloud platform. The result of studying is shown on Figure 2.

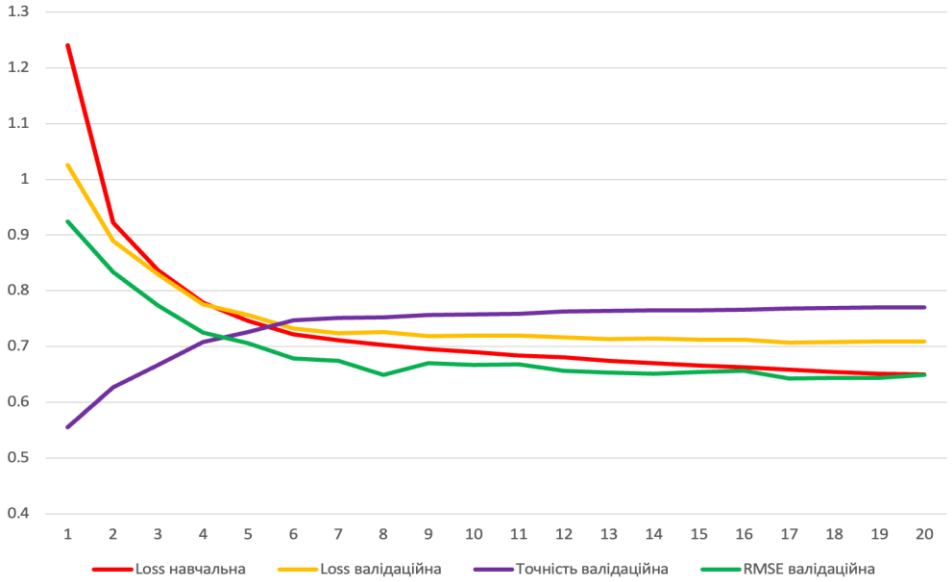


Fig.2. Progress of neural network training

As a result, the accuracy was 76.3%. The value of the logistic loss function metric and the root mean square error were 0.7061 and 0.6478, respectively.

Experiments and researching

The next part of the work is about using the created neural network for improving accuracy of spam detection. A well-known dataset [10] with comments from the youtube.com website was chosen. The dataset itself consists of 1961 reviews, each of which belongs to the spam or non-spam (ham) class. Both classes are represented in the dataset in the same proportion. Just like the previous dataset, this one has also been normalized and vectorized [8, 9]. In addition, all Internet references were selected from the text. To perform tokenization, the TfidfVectorizer function from the sklearn library is used. It analyzes the number of repetitions of each word and leaves only a set of the most popular words. Next, this algorithm assigns a specific vector to each word. The resulting vector representations of words can be used in natural language processing and machine learning. In this case, the simplest way of converting words into a vector was used, when each word is assigned its own vector, which does not depend on other words.

Firstly, the Naive Bayes classifier was first used. A Naive Bayes classifier is a probabilistic classifier based on the Bayes theorem, that based on naive assumptions of independence [11]. The MultinomialNB object from the sklearn library was used for training. As an input, it accepted a vector in which the text is encoded, and as an output it outputs the probability of this text belonging to spam, from 0 to 1. In the result, the accuracy of this method is 88.3%.

In the next step, we add a feature of the sentiment analyze to learning parameters. This feature was calculated using a previously created neural network. As a result, accuracy increased from 88% to 93%. Confusion matrices before and after are shown on Figure 3.

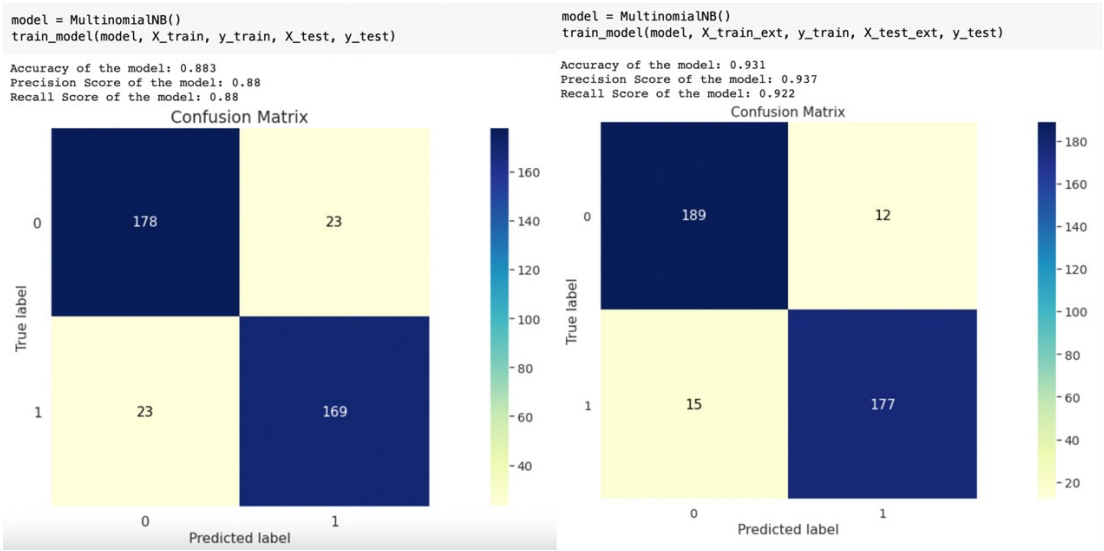


Fig.3. Confusion matrices for Naive Bayes before and after addition a feature of sentiment analysis

The same flow was done for the Random Forest algorithm. Random Forest is a supervised machine learning algorithm used for classification and regression tasks [1]]. It is an ensemble learning method that combines multiple decision trees to make predictions. In a Random Forest model, a set of decision trees is built independently, each using a random subset of the features and a random subset of the training data. Each decision tree is trained to predict the outcome variable based on the subset of features and data it was given. As a result, the accuracy increased from 91% to almost 96%. Confusion matrices before and after are shown on Figure 4.

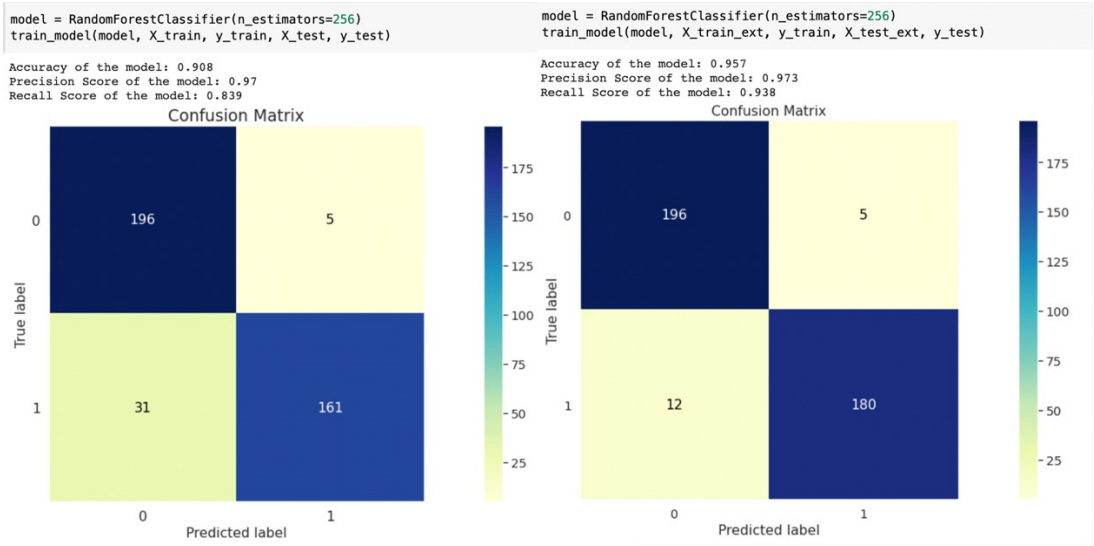


Fig.4. Confusion matrices for Random Forest before and after addition a feature of sentiment analysis

Conclusions

Firstly, we made an analysis of the subject area of systems for analyzing the tone of the text. It showed that there is no universal solution. The opposite situation is for the task of spam detection, for which there are many ready-made systems, but they have a close (private) program code. Therefore, it is not known whether they use the analysis of the tonality of the text or not.

Secondly, we used an LSTM neural network and dataset [7] for its training and validation. Three metrics for evaluating the quality of a neural network were described, and the dataset was analyzed and split into training, validation, and text datasets. The neural network was trained on the Google Colab platform using GPUs. As a result, the neural network was able to evaluate the tone of the comment on a scale from 1 to 5, where the higher the rating, the more emotionally positive the response and vice versa. After training, the neural network achieved an accuracy of 76.3% on the test dataset, and the root mean squared error was 0.6478. This means that error of neural network is less than one class.

Thirdly, we researched how this neural network can improve accuracy of spam detection. In order to do this, a dataset [10] was selected. The text of each response of this dataset was cleaned, tokenized and transformed into a vector in the same way as the previous dataset. When we trained the Naive Bayes classifier algorithm without sentiment analysis, the accuracy was 88.3%, while with the text sentiment analysis the accuracy increased to 93.1%. When we trained the Random Forest without sentiment analysis, the accuracy was 90.8%, while with the text sentiment analysis the accuracy increased to 95.7%.

As a conclusion, the adding feature of sentiment analysis increases the accuracy for both models. The value of the increase in accuracy is 4.8% for the Naive Bayes classifier and 4.9% for the Random Forest. Therefore, sentiment analysis can be used to improve spam detection. It is worth noting that the accuracy of the Random Forest is higher than the accuracy of the Naive Bayes classifier for this task.

References

1. Hopfield J. J. Neural networks and physical systems with emergent collective computational abilities, 1984. С. 147-169.
2. Segaran T. Programming collective intelligence. LA, 2012
3. Deerwester, S. C., Dumais S. T., Landauer T. K. Indexing by Latent Semantic Analysis. 1990. С. 391–407.
4. Gers F. A., Schmidhuber J., Cummins F. Learning to forget: continual prediction with LSTM. UK, 1999. С. 850–855.
5. Goodfellow I., Bengio Y., Courville A. Deep Learning, 2016. 773 с.
6. John E. Kelly I. Steve Hamm Smart Machine, 2014. 147 с.
7. Yelp review full dataset: веб-сайт. URL: http://hidra.lbd.dcc.ufmg.br/datasets/yelp_2015/original/yelp_review_full_csv.tar.gz (дата звернення 01.02.2023).
8. Yang Y., Pedersen J. O. A comparative study of feature selection in text categorization. 1997. С. 412–420.
9. Yang Y., Pedersen J. O. Feature selection in statistical learning of text categorization. – 1997.
10. YouTube Spam Collection Data Set: веб-сайт. URL: https://archive.ics.uci.edu/ml/datasets/Youtube+Spam+Collection (дата звернення 01.02.2023).
11. Alzubi, J., Nayyar, A., Kumar, A. Machine learning from theory to algorithms: an overview, 2018. 43 с.

Oleksandr Iermolaiev **Олександр Єрмолаєв**	Master of Computer Science e-mail: abionics.dev@gmail.com https://orcid.org/0009-0008-1092-2101	Чорноморський національний університет імені Петра Могили
Inessa Kulakovska **Інеса Кулаковська**	PhD of Physical and Mathematical Sciences, Department of Intelligent Information Systems, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine e-mail: Inessa.Kulakovska@chmnu.edu.ua https://orcid.org/0000-0002-8432-1850	Чорноморський національний університет імені Петра Могили

UDC 519.6
https://doi.org/10.31891/csit-2023-1-7

Lesia MOCHURAD, Andrii ILKIV
Lviv Polytechnic National University
Oleksandr KRAVCHENKO
National Technical University «Kharkiv Polytechnic Institute»

A NEW INFORMATION SYSTEM FOR ROAD SURFACE CONDITION CLASSIFICATION USING MACHINE LEARNING METHODS AND PARALLEL CALCULATION

Modern information systems are increasingly used in various areas of our life. One of these is the quality control of the condition of the road surface in order to carry out repair work on time if necessary. The machine learning method can facilitate the control process, which was demonstrated in this work.

Analyzing the road surface condition using image classification requires much pre-classified data and decent computing power. As the modern need for proper quality control of the road surface is high, it is possible to analyze using sensor-recorded data in tabular form and machine learning methods, which should show high accuracy of the classification results. Development and research of an information system for classifying the condition of the road surface were described in this paper, including ways for optimizing similar approaches and improving the results obtained through the use of a greater number of features, in particular, taking into account not only the speed indicators at the given time of the car's movement but also the performance indicators of internal combustion engine. As a result, an information system was developed that classifies the road surface condition using features obtained from various types of sensors and recorded in tabular form. Machine learning methods such as Random Forest, Decision Tree, Support Vector Method, and AutoML library were used to compare accuracy results using a large set of artificial intelligence methods. The best results were obtained using the Random Forest ensemble machine learning method. The analysis of the classifier according to various parameters was carried out, and a search for the best hyperparameters was performed. At the same time, achieving a 91.9% accuracy of road surface condition classification was possible. Parallel calculations were used during model training. As a result, training time was decreased by 5 times with the use of the CPU and by 51 times with the help of the GPU.

Keywords: road condition classification, Random Forest, reference vectors method, decision tree, CUDA technology.

Леся МОЧУРАД, Андрій ІЛЬКІВ
Національний університет «Львівська політехніка»
Олександр КРАВЧЕНКО
Національний технічний університет «Харківський політехнічний інститут»

НОВА ІНФОРМАЦІЙНА СИСТЕМА ДЛЯ КЛАСИФІКАЦІЇ СТАНУ ДОРОЖНЬОГО ПОКРИТТЯ ЗА ДОПОМОГОЮ МЕТОДІВ МАШИННОГО НАВЧАННЯ ТА ПАРАЛЕЛЬНИХ ОБЧИСЛЕНЬ

Аналіз стану дорожнього покриття при класичному підході розпізнавання особливостей із використання зображень потребує великої кількості даних з попередньо підготовленим описом та обчислювальних потужностей. Враховуючи сучасні потреби своєчасного контролю якості покриття дорожніх шляхів, аналіз можливо спростити з використанням показників записаних у табличному виглядів та методів машинного навчання, які у сукупності повинні показати задовільну точність результатів. У роботі було обґрунтовано доцільність розробки та дослідження інформаційної системи класифікації стану дорожнього покриття, а також визначено як ключовий напрямок оптимізацію аналогічних підходів та покращення отриманих результатів шляхом використання більшої кількості ознак, зокрема врахування не лише швидкісних показників в момент руху автомобіля, але й показників роботи двигуна внутрішнього згоряння, яке напряму пов'язана з можливістю здійснювати рух. У результаті розроблено інформаційної системи, що класифікує стан дорожнього покриття за ознаками отриманими з різного виду датчиків та записаних у табличному вигляді. Використано методи машинного навчання такі як Випадковий Ліс, Дерево Рішень, Метод опорних векторів та бібліотеку AutoML, яка дозволила провести порівняння точності результатів з використанням великого набору методів штучного інтелекту. Найкращі результати вдалось отримати за допомогою ансамблевого методу машинного навчання Rendom Forest. Проведено аналіз роботи класифікатора при різних параметрах та виконано пошук найкращих гіперпараметрів. При цьому вдалось досягти точності класифікації стану дорожнього покриття рівній 91.9 %. Для можливості прискорення отримання рішення було застосовано паралельні обчислення при тренуванні моделі. В результаті отримано показник прискорення у 5 разів з використанням CPU та у 51 раз з використанням GPU.

Ключові слова: класифікація стану дорожнього покриття, випадковий ліс, метод опорних векторів, дерево рішень, технологія CUDA.

Introduction

When a car passes a surface of different types and conditions, changes in the speed, trajectory, and smoothness of the vehicle's movement occur. Thanks to the use of motion sensors, and sensors of the internal combustion engine, it is possible to classify the type of road surface, driving style, and road traffic [1]. Target variable classification [2] based on data gathered during the vehicle movement allows to combine all of this into effective approach for fast classification to determine the condition of the road surface at a certain moment of the time of operation of the sensors and additionally geolocation of the vehicle at the same moment.

There are two popular approaches to classifying the condition of a road surface:

- Use of tabular data obtained while driving the car;
- Use of overlay images.

Road surface condition classification using images will guarantee more accurate predictions [3] but requires a large amount of image data. Collecting and forming a suitable dataset is much easier in case of classification based on the use of tabular data rather than image-based classification. The approach of data collection using the smartphone gyroscope is popular [4–6]. However, the advantage of the developed information system will be the use of tabular data obtained from the accelerometer and the electronic control unit of the car's internal combustion engine in combination with classifiers, the best of which will be selected using the AutoML library for the high-level Python programming language and optimized using parallel computing algorithms [7]. The use of several existing machine learning algorithms to solve the given task will provide an opportunity to conduct a comparative analysis with the ability to determine the most optimal approach.

Therefore, the purpose of this paper is to create an information system for classifying the condition of the road surface using input data obtained from sensors installed in the car based on machine learning algorithms, improving the results of the system by using parallel computing and conducting a comparative analysis of the results of the system when applying hyperparameter settings classifiers.

The object of the study is the information system of classification or prediction of the condition of the road surface based on tabular data using machine learning algorithms.

The subject of the research will be the use of machine learning methods such as Random Forest, Decision Tree, Support Vector Method, and the AutoML library, which will allow comparison of the accuracy of results using a large set of artificial intelligence methods.

A large number of publications and studies in the field of road pavement condition classification confirm the relevance of the chosen topic [8, 9]. Since pavement maintenance is an important task for maintaining the stable operation of urban infrastructure, automating the identification of areas that require immediate maintenance or are in unsatisfactory condition will facilitate the appropriate analysis.

Related works

During the analysis of related works, modern approaches to solving the problem of road surface condition classification and the optimal application of machine learning methods to solve this problem were considered.

The solution to the problem of road surface condition classification using classifiers are described in [10]. In this work, a study was carried out, the purpose of which was to determine the features from the data set that most affect the accuracy of the obtained results and to carry out classification using the analysis of vehicle vibrations. When analyzing the input data, there is a high probability of anomalies due to the sensitivity of the data collection sensors. Carrying out a sensitivity analysis of the received readings will make it possible to increase the weight of some features compared to others based on their value when obtaining classification results. Using the analysis of the time series of car oscillations, the authors determined the average deviation of the signs. They conducted an analysis based on frequency oscillations from the corresponding sensors. The impact of various features used in classification on the accuracy of the selected classifiers RF, DT, and MLP was also investigated. During the analysis, two sets of data were used: data obtained from sensors and simulated data. The best accuracy of 87% was obtained precisely for the simulation data set, which may indicate a linear relationship between features and affect accuracy.

A similar problem was considered by the paper's authors [4]. The study of road surface anomalies is focused on the data obtained from the smartphone's accelerometer, gyroscope, and GPS data. The authors of this scientific work suggest using a smartphone when analyzing the road surface to determine its condition. Indicators of changes in acceleration and vibration of the smartphone were analyzed, which made it possible to classify each part of the road surface with the corresponding geolocation. The results were obtained using machine learning algorithms and a multilayer neural network. At the same time, the authors tried to combine the results of the smartphone's accelerometer and gyroscope with the obtained road surface images. Since assigning labels to each image in order to further use them in a complex classification requires a large amount of human labor, the authors investigated the possibility of automatic classification and assignment of labels to images for further work with convolutional neural networks. The disadvantage of this work is that with a data set of 1010 rows, it takes 70 seconds to train the model with an accuracy of 88% when using the SVM and RF algorithms.

In paper [11], the authors tried to analyze the state of the road surface and search for its anomalies using a geoinformation system. Only Ox, Oy, and Oz acceleration indicators were used for the analysis. The classification of the condition of the road surface was carried out by an impulse neural network. As an aproach was chosen the ensemble learning of the classification model for the simultaneous determination of the condition of both roads with and without asphalt. The result of the work of the geoinformation system proposed in this article was 99.9% in determining the type of road path and 99.8% during the classification of the state of this path. Since the training dataset used a dataset of 2835 rows to train the model for pavement type detection and a dataset of 4300 rows to train the pavement condition classification model, the model could have been overtrained. If we compare the results with the work [10], then the accuracy of predictions should differ by a factor of two between actual and simulated data.

The optimization of execution time by using algorithms of parallel calculations was considered by the authors of the work [12]. The authors proposed to speed up the training of the RF model by performing parallel training on four threads simultaneously. As a result, it was experimentally proven that execution on four threads and with a given parameter of the number of trees equal to 100 training took 77 seconds, which is twice as fast as training using only two threads. However, work [2] shows that using GPU for parallelization is 83.4 times faster than parallelization using 8 CPU threads.

Improving the running time of model training with GPU was considered in [13]. Even modern equipment can only sometimes cope with the tasks when working with large data sets. The paper proposes an approach that includes the application of a reduction algorithm based on the stochastic distribution of neighbors in the data set in combination with a new approach to calculating the KNN graph using GPU. This approach showed a speedup of 460%.

The advantage of our proposed approach over the methods mentioned above will be taking into account not only the speed indicators at the time of the car's movement but also the performance indicators of the internal combustion engine, which are directly related to the ability to move. Also, using parallel computing algorithms will provide an opportunity to obtain classification results faster without losing the accuracy of these predictions.

Methodology

For achieving aim of this paper next items should be covered:

- preparation for data set classification;
- classification using parallel computing algorithms;
- visualize and compare the obtained results.

Classification of the road surface condition will be done using the data set [14], which consists of 17 features and 24,957 records. The following 17 features are divided into categories depending on the type of sensor from which the data was received:

Data obtained from the accelerometer:

- AltitudeVariation – change in height during movement
- VerticalAcceleration – vertical acceleration of the car
- LongitudinalAcceleration – horizontal acceleration of the car

Data obtained from the car's motion sensors:

- VehicleSpeedAverage – the average speed of the car
- VehicleSpeedVariance – the difference between the speed of the car and speed limit
- VehicleSpeedVariation – average change in vehicle speed
- VehicleSpeedInstantaneous – the speed of the car at a given moment of time

Data obtained from the car's electronic control unit:

- EngineLoad
- EngineCoolantTemperature – engine cooling system temperature
- ManifoldAbsolutePressure – absolute pressure in the exhaust manifold
- EngineRPM – engine revolutions per minute
- MassAirFlow – air flow mass
- IntakeAirTemperature – air temperature at the inlet to the fuel combustion valves
- FuelConsumptionAverage – average fuel consumption by car

Target features:

- roadSurface – condition of the road surface
- traffic – car traffic at a given moment of time
- drivingStyle – the style of driving a car at a given moment in time

As a result of the classification by several target variables, it is possible to achieve a higher accuracy of the classification since the performance indicators of the engine will be used for the analysis, which, depending on traffic jams or aggressive driving style, will increase the load. If these factors are taken into account during classification, accuracy of machine learning algorithms results will be increased [15].

When preparing the data set, possible anomalies that can affect the results' quality must be considered. Such anomalies can be both missing values and values that strongly deviate from the average. To eliminate this problem, we need to normalize the data set. The z-score and min-max methods will be used [16]. RF, SVM, and DT classifiers from the sklearn library for the high-level Python programming language will be the main tools for classification. The accuracy and performance of the trained models will be evaluated using F1-Score, Precision-Recall, MSE, and ROC-AUC metrics. To analyze the obtained results, the matplotlib library will be used to visualize the results and present them as graphs of functions. To optimize the training time of the classification models, training using multiple CPU and GPU threads will be applied to the algorithm that showed the highest accuracy – RF.

Ensemble machine learning methods will be applied to solve the problem of pavement condition classification, such as the RF regression method, which is based on the regression construction of numerous decision

trees during model training, and the SVM method, which performs classification by building models called support vector networks [17].

When training an RF model, the training sample will be divided into random subsamples with replication. In this case, some records may enter the sample several times, and others not. The next step will be to randomly select several features, after which the construction of decision trees will begin, which will classify this subsample and only on the part of the randomly selected features. The most valuable features are selected using the Gini criterion.

In the case of training the RF model on multiple streams simultaneously, the bootstrap aggregation will be applied. Since the sample will be divided during training, the summation of the nearest neighboring classifiers will allow for obtaining a model at the end, which will be a collection of classifiers performed on separate streams[18].

Trees are built until the entire subsample is passed. Thus, not only the accuracy of the classification will depend on the number of built trees, but also the time required for training the model will increase. The optimal number of trees is selected in such a way as to minimize the error of the classifier during the analysis.

An important step in training the RF model is to evaluate the importance of each feature of the training sample. In order to assess the importance of each sample parameter, a model is trained on this set with error estimation and parameter shuffling at the end of each iteration. The importance of the parameter is estimated by calculating the errors before and after the shuffle. Error-values are normalized, taking into account the standard deviation.

To evaluate the performance of classifiers, such metrics as:

- F1 score,
- ROC AUC,
- Precision Recall,
- Accuracy.

Here, accuracy is a percentage value that indicates the number of correctly classified target features. It is calculated according to the following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP – positive samples that are correctly classified,

FN – negative samples that are incorrectly classified,

FP – positive samples that are incorrectly classified,

TN – negative samples that are correctly classified.

To measure the rate of acceleration and efficiency during the parallel execution of the classification model's training process, the classifier's computational complexity needs to be determined. For RF, it is calculated by the formula: $O(n * \log(n) * d * k,$

Where n – the number of records in the data set;

k – the number of trees during training;

d – the size of the training data set.

In order to calculate the value of the theoretical acceleration indicator, it is necessary to find the ratio of computational complexity when performing training consecutively and in parallel according to the formula:

$$S_p(n) = \frac{T_1(n)}{T_p(n)},$$

where p – number of threads;

$T_1(n)$ – time complexity of sequential execution of the algorithm;

$T_p(n)$ – time complexity of parallel execution of the algorithm for p threads.

The efficiency indicator is calculated according to the formula: $E_p(n) = \frac{S_p(n)}{p} = \frac{O(n*\log(n)*d*k}{p*O(\frac{n}{p}*\log(\frac{n}{p})*d*k)}$.

The performance indicator cannot be calculated for training performed on the GPU because the GPU has a complex structure of cores called pipelines.

Experiments

The training of classification models was carried out on a machine with an Intel core I7-6700 CPU with 4 cores and 8 threads. NVIDIA GT970M video card with 2 gigabytes of video memory was used to train the model using the GPU.

The pavement condition classification study was conducted on the previously described data set split 80 : 20 into training and test data sets.

The definition of the classifier with the highest accuracy was carried out using the AutoML software library. As an input, it receives a training data set, after which it analyzes features, their interdependence, and the type of the target variable. The result of the work of the AutoML library is a rating table of the accuracy of the obtained classifiers,

which were determined to be the most suitable for classification. Each classifier trained the model with standard parameters. Fig. 1 shows the rating table of classifiers.

model_id	rank	ensemble_weight	type	cost	duration
2	1	0.52	random_forest	0.066227	44.534895
3	2	0.08	extra_trees	0.122478	34.161803
8	3	0.08	mlp	0.148524	51.222872
11	4	0.06	mlp	0.277987	31.608544
10	5	0.16	multinomial_nb	0.409596	2.396460
7	6	0.10	bernoulli_nb	0.545404	2.334953

Fig. 1 Rating table of classification quality of trained models

The RF classifier showed the best accuracy, with an accuracy of 85%, training the model in 44.5 seconds. Training and searching for the best model among the available ones took 6 minutes. The following metric results were obtained (see Fig. 2):

	precision	recall	f1-score	support
FullOfHolesCondition	0.98	0.96	0.97	636
SmoothCondition	0.99	0.98	0.99	3012
UnevenCondition	0.99	0.94	0.97	1344
HighCongestionCondition	1.00	0.89	0.94	647
LowCongestionCondition	0.99	0.99	0.99	3701
NormalCongestionCondition	0.98	0.91	0.95	644
AggressiveStyle	0.87	0.48	0.61	555
EvenPaceStyle	0.95	0.98	0.97	4437
micro avg	0.98	0.95	0.96	14976
macro avg	0.97	0.89	0.92	14976
weighted avg	0.97	0.95	0.96	14976
samples avg	0.98	0.95	0.96	14976

Fig. 2 Performance metrics of the best RF classifier model

Since the model was trained using the default value of the number of trees, depth, and the minimum number of tree leaves, a high accuracy value could not be achieved. To improve the results, it is necessary to select hyperparameters that could increase the accuracy of the classifier. The RandomizedSearchCV function from the sklearn library will be used for this.

```
[Parallel(n_jobs=-1)]: Done 150 out of 150 | elapsed: 56.2min finished
{'n_estimators': 307, 'min_samples_split': 20, 'min_samples_leaf': 4,
 'max_features': 'sqrt', 'max_depth': 90, 'criterion': 'gini', 'bootstrap': True}
Time to get best hyperparameters = 3394.156 seconds
```

Fig. 3 The result of the hyperparameter selection function

As we can see from Fig 3, the applied function selected the most suitable among many possible parameters, so let us build a new tree using these hyperparameters.

	precision	recall	f1-score	support
FullOfHolesCondition	0.99	0.97	0.98	632
SmoothCondition	0.99	0.99	0.99	3062
UnevenCondition	0.99	0.98	0.98	1298
HighCongestionCondition	0.99	0.96	0.97	588
LowCongestionCondition	0.99	1.00	0.99	3781
NormalCongestionCondition	1.00	0.94	0.97	623
AggressiveStyle	0.92	0.58	0.71	589
EvenPaceStyle	0.95	0.99	0.97	4403
micro avg	0.98	0.97	0.97	14976
macro avg	0.98	0.93	0.95	14976
weighted avg	0.98	0.97	0.97	14976
samples avg	0.98	0.97	0.97	14976

```
Model training time is  252.8594033718109
Model Classification Accuracy =  0.9186698717948718
```

Fig. 4 Results of RF classifier training using the obtained parameters

The training result is a model with a classification accuracy of 91.8%, which is a satisfactory accuracy when using the classifier. However, the time required to train the RF model is 252 seconds (see Fig. 4).

When using this classifier to train a model using even more data, the time will increase, so to optimize this approach. It is necessary to apply parallelization and compare the obtained results. In order to solve this problem and reduce the training time of the classification model, an approach using parallel training using the different number of threads and computing technology using GPU - CUDA will be applied.

When training the RF model for parallel computation, the training data set is divided into number particles, and a classification tree is built for each. The simultaneous training of several trees and combining their results makes obtaining a trained model faster [18].

FullOfHolesCondition	0.98	0.98	0.98	627
SmoothCondition	1.00	0.99	0.99	3025
UnevenCondition	1.00	0.98	0.99	1340
HighCongestionCondition	1.00	0.95	0.97	637
LowCongestionCondition	0.99	1.00	0.99	3716
NormalCongestionCondition	1.00	0.94	0.97	639
AggressiveStyle	0.89	0.58	0.71	540
EvenPaceStyle	0.95	0.99	0.97	4452
micro avg	0.98	0.97	0.97	14976
macro avg	0.98	0.92	0.95	14976
weighted avg	0.98	0.97	0.97	14976
samples avg	0.98	0.97	0.97	14976

```
Model training time is  121.47707033157349
Model Classification Accuracy =  0.9192708333333334
```

Fig. 5 RF model training on 2 threads

After training the model using two streams (see Fig. 5), it was possible to reduce the time of the training process by 2.1 times compared to sequential training.

	precision	recall	f1-score	support
FullOfHolesCondition	0.98	0.98	0.98	627
SmoothCondition	1.00	0.99	0.99	3025
UnevenCondition	1.00	0.98	0.99	1340
HighCongestionCondition	1.00	0.95	0.97	637
LowCongestionCondition	0.99	1.00	0.99	3716
NormalCongestionCondition	1.00	0.94	0.97	639
AggressiveStyle	0.89	0.58	0.71	540
EvenPaceStyle	0.95	0.99	0.97	4452
micro avg	0.98	0.97	0.97	14976
macro avg	0.98	0.92	0.95	14976
weighted avg	0.98	0.97	0.97	14976
samples avg	0.98	0.97	0.97	14976

```
Model training time is  77.025221824646
Model Classification Accuracy =  0.9192708333333334
```

Fig. 6 RF model training on 4 threads

When training the classification model using 4 streams (see Fig. 6), it took 3.2 times less time compared to sequential execution.

	precision	recall	f1-score	support
FullOfHolesCondition	0.98	0.98	0.98	627
SmoothCondition	1.00	0.99	0.99	3025
UnevenCondition	1.00	0.98	0.99	1340
HighCongestionCondition	1.00	0.95	0.97	637
LowCongestionCondition	0.99	1.00	0.99	3716
NormalCongestionCondition	1.00	0.94	0.97	639
AggressiveStyle	0.89	0.58	0.71	540
EvenPaceStyle	0.95	0.99	0.97	4452
micro avg	0.98	0.97	0.97	14976
macro avg	0.98	0.92	0.95	14976
weighted avg	0.98	0.97	0.97	14976
samples avg	0.98	0.97	0.97	14976

```
Model training time is  55.0640594959259
Model Classification Accuracy =  0.9192708333333334
```

Fig. 7 RF model training on 8 threads

When using the maximum possible number of streams (8 threads, see Fig. 7) of the machine on which the classifier's training was performed, it was possible to reduce the time by 5 times. Notably, the accuracy of such a model was the same regardless of the number of threads.

Cuml software library was used to train the classifier model using GPU and CUDA technology. Although training a classification model using a GPU requires preprocessing parts of the training data set and loading them directly onto the GPU, as seen in Fig. 8, the training process is much faster.

```
                    FullOfHolesCondition       0.99      0.98      0.98       627
                        SmoothCondition       1.00      0.99      0.99      3025
                       UnevenCondition        1.00      0.98      0.99      1340
                  HighCongestionCondition     1.00      0.95      0.97       637
                   LowCongestionCondition     0.99      1.00      0.99      3716
                NormalCongestionCondition     0.99      0.93      0.96       639
                      AggressiveStyle         0.90      0.58      0.70       540
                       EvenPaceStyle          0.95      0.99      0.97      4452

                           micro avg          0.98      0.97      0.97     14976
                           macro avg          0.98      0.92      0.95     14976
                        weighted avg          0.98      0.97      0.97     14976
                         samples avg          0.98      0.97      0.97     14976

Model training time is 5.8234123
Model Classification Accuracy =  0.9184695512820513
```

Fig. 8 Training the RF model using the GPU

The time required for training was reduced by 51 times, indicating CUDA technology's effectiveness in parallel training of the classifier model.

We will consider the results of the training, namely the execution time and accuracy of the trained classification model, using the parallel implementation of the program on CPU and GPU, and conduct a comparative analysis.

Table 1

Program execution time with sequential and parallel training, s

Sequential execution	Parallel execution, number of threads			
	2	4	8	GPU
252.85	121.47	77.02	55.06	5.82

As we can see from Table 1, although the training time of the model with the maximum possible number of parallel threads for the architecture we use is 5 times less than the training time with sequential processing, it is pretty slow compared to training on the GPU. This indicates the incredible computing power of the video card and the advantages of performing parallel calculations on it relative to calculations on the processor.

Table 2

Accuracy of the trained model, %

Sequential execution	Parallel execution, number of threads			
	2	4	8	GPU
91.86	91.92	91.92	91.92	91.84

From the results presented in Table 2, the accuracy experienced slight deviations depending on the number of threads or the computational unit on which the training was performed. However, the quality of the performance was not affected.

Now let us calculate experimental indicators of acceleration and efficiency of parallel algorithms at different numbers of threads if parallel calculations are carried out on the processor, as well as indicators of acceleration of parallel algorithms for GPU.

First, let us perform a theoretical speedup evaluation for different numbers of parallel threads that we will use to train our tree. It should be emphasized that it is analytically impossible to calculate a theoretical estimate of the acceleration of model training on GPU since the number of graphics cores used during training is still being determined.

$$S_2(24957) = \frac{O(24957 * log(24957) * 17 * 800}{O(\frac{24957}{2} * log\left(\frac{24957}{2}\right) * 17 * 800)} \approx 2.1;$$

$$S_4(24957) = \frac{O(24957 * log(24957) * 17 * 800}{O(\frac{24957}{4} * log\left(\frac{24957}{4}\right) * 17 * 800)} \approx 3.6;$$

$$S_8(24957) = \frac{O(24957 * log(24957) * 17 * 800}{O(\frac{24957}{8} * log\left(\frac{24957}{8}\right) * 17 * 800)} \approx 5.8.$$

Table 3

The acceleration indicators of the parallel algorithm are obtained based on numerical experiments

Number of threads			GPU
2	**4**	**8**	
2.1	3.2	5	51

Table 3 shows the acceleration indicators obtained with the help of parallel execution of the algorithm for the different number of threads when working on the CPU and with the help of execution on the GPU.

The following conclusion can be drawn by comparing the theoretical estimation of acceleration and the results obtained in Table 3. The training was conducted on a laptop with a 4-core processor that supports Hyper-Threading technology, which provides an opportunity to obtain additional 4 computing threads (virtual threads). Although we can use 8 parallel threads for training, the load on the processor increases significantly, which leads to a loss of efficiency in the calculation of operations. Therefore, when increasing the number of parallel threads we use to train the model, we can see that the actual speedup estimate is significantly different from the theoretical one. If we talk about the parallel execution of training on the processor, then when the number of threads increases, the acceleration value increases. However, as we can see, compared to the execution of parallel calculations on the video card, the acceleration obtained by execution on the processor is much smaller, which once again indicates the incredible power of the execution of calculations with the help of the video card.

Since a theoretical acceleration estimate was made, which is the maximum we can achieve, we assume that the theoretical efficiency estimate is also the maximum possible and being equal to 1.

Table 4

Actual performance indicators of the parallel algorithm with different number of threads of the processor

Number of threads		
2	4	8
1	0.8	0.625

Analyzing Table 4, we can see that the efficiency decreases as the number of parallel threads we use to train the model increases. This can be explained by the fact that when the number of parallel threads we use for calculation increases, the load on the processor increases, which in turn, causes the calculation to be less efficient. Therefore, it will be advisable to use the GPU as a more efficient unit when performing calculations.

Conclusions

Modern information systems are increasingly used in various areas of our life. One of these is the quality control of the condition of the road surface in order to carry out repair work on time if necessary. The machine learning method can facilitate the control process, which was demonstrated in this work.

Data analysis methods for classifying road surface conditions were considered in detail. The research used a data set that contains approximately 25,000 records with 17 parameters about the car's movement, engine operation, road traffic conditions, driver behavior, and road surface conditions.

During the research, several models were successfully trained for the classification of the road surface condition, and a comparative analysis of their work was carried out using the AutoML software library.

The results, which were the best obtained using the RF ensemble method of machine learning, were discussed in more detail. After analyzing the performance of the classifier with different parameters and searching for the best hyperparameters, it was possible to achieve an accuracy of 91.9% classification of the road surface condition.

To improve the performance of the model, parallel computing was applied when training the model, which made it possible to speed up the training by 5 times using the CPU and 51 times using the GPU.

In this way, information technology was developed to train a classification model based on an input data set in the shortest possible time. This model can further be used to analyze and classify the data set for automated determination of the condition of the road surface.

References

1. Menegazzo Jeferson, Von Wangenheim Aldo. Road Surface Type Classification Based on Inertial Sensors and Machine Learning: A Comparison Between Classical and Deep Machine Learning Approaches for Multi-Contextual Real-World Scenarios. *Computing* 103(4), 2021. doi:10.1007/s00607-021-00914-0.
2. Mochurad L., Ilkiv A. A novel method of medical classification using parallelization algorithms. *Comput. Syst. Inf. Technol.*, № 1, pp. 23–31, Apr. 2022, doi: 10.31891/CSIT-2022-1-3.
3. Li J., Liu T., Wang X., and Yu J. Automated asphalt pavement damage rate detection based on optimized GA-CNN. *Autom. Constr.*, Vol. 136, p. 104180, Apr. 2022, doi: 10.1016/j.autcon.2022.104180.
4. Basavaraju A., Du J., Zhou F., and Ji J. A Machine Learning Approach to Road Surface Anomaly Assessment Using Smartphone Sensors. *IEEE Sens. J.*, 20(5), pp. 2635–2647, 2020, doi: 10.1109/JSEN.2019.2952857.
5. Setiawan B.D., Serdult U.I. and Kryssanov V. Smartphone Sensor Data Augmentation for Automatic Road Surface Assessment Using a Small Training Dataset, in *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Jeju Island, Korea (South), Jan. 2021, pp. 239–245. doi:10.1109/BigComp51126.2021.00052.
6. Wu C. *et al.* An Automated Machine-Learning Approach for Road Pothole Detection Using Smartphone Sensor Data, *Sensors*, 20(19), p. 5564, Sep. 2020, doi: 10.3390/s20195564.

7. Mochurad L., Boyko N. Solving Systems of Nonlinear Equations on Multi-core Processors. *Advances in Intelligent Systems and Computing IV*, Vol. 1080, N. Shakhovska and M. O. Medykovskyy, Eds. Cham: Springer International Publishing, 2020, pp. 90–106. doi: 10.1007/978-3-030-33695-0_8.
8. Martinez-Ríos E.A., Bustamante-Bello M.R., Arce-Sáenz L.A. A Review of Road Surface Anomaly Detection and Classification Systems Based on Vibration-Based Techniques. *Applied Sciences*. 2022; 12(19):9413. doi:10.3390/app12199413.
9. Nausheen Saeed, Roger G. Nyberg & Moudud Alam. Gravel road classification based on loose gravel using transfer learning, *International Journal of Pavement Engineering* 2022*,* doi: 10.1080/10298436.2022.2138879.
10. Ferjani I., Ali Alsaif S. How to get best predictions for road monitoring using machine learning techniques. *PeerJ Computer Science* 8:e941, 2022, doi:10.7717/peerj-cs.941.
11. Agebure M.A., Oyetunji E.O., Baagyere E.Y. A three-tier road condition classification system using a spiking neural network model. *J. King Saud Univ. - Comput. Inf. Sci.*, 34(5), pp. 1718–1729, May 2022, doi: 10.1016/j.jksuci.2020.08.012.
12. Azizah N., Riza L.S., Wihardi Y. Implementation of random forest algorithm with parallel computing in R. *J. Phys. Conf. Ser.*, Vol. 1280, p. 022028, Nov. 2019, doi: 10.1088/1742-6596/1280/2/022028.
13. Meyer B.H., Pozo A.T.R., Zola W.M.N. Improving Barnes-Hut t-SNE Algorithm in Modern GPU Architectures with Random Forest KNN and Simulated Wide-Warp. *ACM J. Emerg. Technol. Comput. Syst.*, 17(4), pp. 1–26, Jun. 2021, doi: 10.1145/3447779.
14. A set of data about car movement information, the operation of its internal combustion engine. URL: https://www.kaggle.com/code/absolutegaming/road-prediction/data.
15. Silva N., Soares J., Shah V., Santos M.Y., Rodrigues H. Anomaly Detection in Roads with a Data Mining Approach. *Procedia Comput. Sci.*, Vol. 121, pp. 415–422, 2017, doi: 10.1016/j.procs.2017.11.056.
16. Kappal S. Data normalization using median median absolute deviation MMAD based Z-score for robust predictions vs. min–max normalization. *Lond. J. Res. Sci. Nat. Form.*, 19(4): 39-44, 2019.
17. Christine Dewi. Random Forest and Support Vector Machine on Features Selection for Regression Analysis. *International journal of innovative computing, information & control: IJICIC* 15(6):2027–2037, 2019, doi: 10.24507/ijicic.15.06.2027.
18. Bruce P.C. and Bruce A. Practical statistics for data scientists: 50 essential concepts, *First edition. Sebastopol, CA: O'Reilly*, 2017.
19. Lindroth L. Parallelization of Online Random Forest. 2021. Accessed: Jun. 05, 2022. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:bth-21098.

Lesia Mochurad **Леся Мочурад**	PhD, Associate Professor of Department of Artificial Intelligence, Lviv Polytechnic National University, Lviv, Ukraine e-mail: lesia.i.mochurad@lpnu.ua. https://orcid.org/0000-0002-4957-1512	кандидат технічних наук, доцент, доцент кафедри систем штучного інтелекту національного університету "Львівська політехніка", Львів, Україна
Andrii Ilkiv **Андрій Ільків**	student of Department of Artificial Intelligence, Lviv Polytechnic National University, Lviv, Ukraine e-mail: andrii.ilkiv.knm.2018@lpnu.ua https://orcid.org/0000-0001-6438-0784	студент кафедри систем штучного інтелекту національного університету "Львівська політехніка", Львів, Україна
Oleksandr Kravchenko **Олександр Кравченко**	PhD student of National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine e-mail: askraff@gmail.com https://orcid.org/0000-0002-6169-1250	аспірант 2-го курсу (Комп'ютерні науки), Національний технічний університет "Харківський політехнічний інститут", Харків, Україна

UDC 004
https://doi.org/10.31891/csit-2023-1-8

Tetiana OKHRIMENKO, Serhii DOROZHYNSKYI, Bohdan HORBAKHA
National aviation university, Kyiv, Ukraine

ANALYSIS OF QUANTUM SECURE DIRECT COMMUNICATION PROTOCOLS

The development of modern computer technologies endangers the confidentiality of information, which is usually ensured by traditional cryptographic means. This circumstance forces us to look for new methods of protection. In view of modern trends, quantum cryptography methods can become such alternatives, which allow solving a number of important cryptographic problems, for which the impossibility of solving using only classical (that is, non-quantum) communication has been proven. Quantum cryptography is a branch of quantum informatics that studies methods of protecting information by using quantum carriers. The possibility of such protection is ensured by the fundamental laws of quantum mechanics. One of the promising directions of quantum cryptography is Quantum Secure Direct Communication (QSDC) that offers secure communication without any shared key. A characteristic feature of this method is the absence of cryptographic transformations, accordingly, there is no key distribution problem. The purpose of this work is a general overview of quantum cryptography protocols, finding their weak points for further development and improvement, as well as identifying vulnerabilities to different attacks.

The article analyzes new methods and protocols, as well as presents their advantages and disadvantages. Based on partial generalizations of theoretical provisions and practical achievements in the field of quantum cryptography, a generalized classification was developed. By comparing various factors of the protocols, and their resistance to certain cyberattacks, we have the opportunity to identify several problems in this field and expand the possibilities for choosing appropriate methods for building modern quantum information protection systems. In accordance with this, conclusions were presented regarding the use of protocols and increasing the level of their effectiveness.

Keywords: quantum cryptography, classification, quantum direct secure communication, quantum key distribution.

Тетяна ОХРІМЕНКО, Сергій ДОРОЖИНСЬКИЙ, Богдан ГОРБАХА
Національний авіаційний університет, Київ, Україна

АНАЛІЗ ПРОТОКОЛІВ КВАНТОВОГО ПРЯМОГО БЕЗПЕЧНОГО ЗВ'ЯЗКУ

Розвиток сучасних обчислювальних технологій ставить під загрозу конфіденційність інформації, що майже завжди забезпечується традиційними криптографічними засобами. Ця обставина змушує шукати нові методи захисту. З огляду на сучасні тенденції, такими альтернативами можуть стати методи квантової криптографії, що дозволяють вирішити немало складних завдань, які неможливо виконати за неквантового обміну інформацією. Квантова криптографія - розділ квантової інформатики, що вивчає методи захисту інформації за допомогою квантових носіїв. Можливість такого захисту забезпечується фундаментальними законами квантової механіки. Одним з перспективних напрямків квантової криптографії є квантовий захищений прямий зв'язок (Quantum Secure Direct Communication, QSDC), який забезпечує безпечний зв'язок без спільного ключа. Характерною особливістю цього методу є відсутність криптографічних перетворень, відповідно, відсутня проблема розподілу ключів. Метою даної роботи є загальний огляд протоколів квантової криптографії, пошук їх слабких місць для подальшого розвитку та вдосконалення, а також виявлення вразливостей до різних атак.

У статті проведено аналіз нових методів та протоколів, а також представлено їх переваги та недоліки. На основі часткових узагальнень теоретичних положень та практичних досягнень у галузі квантової криптографії було розроблено узагальнену класифікацію. Порівнюючи різні фактори протоколів та їх стійкість до певних кібератак, ми маємо можливість виявити ряд проблем у цій галузі та розширити можливості вибору відповідних методів для побудови сучасних систем квантового захисту інформації. У відповідності до цього, представлено висновки щодо використання протоколів та підвищення рівня їх ефективності.

Ключові слова: квантова криптографія, класифікація, квантовий прямий безпечний зв'язок, квантовий розподіл ключів.

Introduction

Quantum cryptography is a branch of quantum informatics that studies methods of protecting information by using quantum carriers. The possibility of such protection is ensured by the fundamental laws of quantum mechanics. One of the promising directions of quantum cryptography is Quantum Secure Direct Communication (QSDC) that offers secure communication without any shared key. A characteristic feature of this method is the absence of cryptographic transformations, accordingly, there is no key distribution problem.

Types of QSDC protocols [1]: 1) Ping-Pong (PP) protocol (various variants) also known as deterministic protocol, 2) protocols with block transmission of entangled qubits, 3) protocols with single qubits, and 4) protocols with groups of entangled qubits. Most of the QSDC protocols proposed so far require the transfer of qubits in blocks. This makes it possible to detect the eavesdropping of the quantum channel before the transmission of the message itself and in this way guarantee the security of the transmission – if the eavesdropping is detected before the transmission of the message, then the legitimate parties interrupt the session and no information is leaked to the attacker. However, to store such blocks of qubits, a large quantum memory is required. Quantum memory technology is actively being developed, but it is still far from mass application in standard telecommunications equipment. Therefore, from the point of view of technical implementation, protocols in which transmission is carried out by single qubits or small groups of them (per one cycle of the protocol) are preferred. Few such protocols have been proposed, and they have only asymptotic security, that is, the attack will be detected with a high probability, but before that, the

attacker will be able to receive some part of the message. Accordingly, there is a problem in strengthening the security of such protocols, that is, creating such methods of preprocessing the transmitted information that will make the information intercepted by the attacker useless for him.

The **purpose of this work** is a general overview of quantum cryptography protocols, finding their weak points for further development and improvement, as well as identifying vulnerabilities to different attacks.

Main part

In QSDC protocols, Alice encodes a secret message consisting of several qubits using a pre-selected encoding rule and sends them to Bob [2]. After some security checks, the recipient can accept the secret message. If the protocol is designed incorrectly, it can give Eve a chance to impersonate a legitimate party. To avoid this, legitimate parties must verify the legitimacy of other parties, which is required by quantum authentication protocols.

In 2006 was introduced first QSDC protocol with authentication, and many researchers have worked in this field since then. Several quantum cryptography protocols have been proven to be invulnerable to various common attacks such as intercept and replay attacks, impersonation attacks, denial of service attacks, man-in-the-middle attacks, Trojan horse attacks, etc. These are active attacks, meaning an interceptor can access the qubits being transmitted in the quantum channel between legitimate parties and actively participate in the protocol. Some passive attacks can also cause information leakage problems in communication protocols.

In 2020, the QSDC protocol, which works on the principle of combining a single photon and a pair EPR was introduced and mutual authentication was achieved. For simplicity, it is called the YZCSS protocol (Yan, Zhang, Chang, Sun, Sheng protocol) [3]. In this protocol, Alice, the sender of the message, prepares pairs of qubits corresponding to the secret message and its authentication identifier. She sends all the qubits to Bob, the recipient of the message, and Bob uses its authentication keys to recover the secret data. However, the YZCSS protocol is not immune to interception and retransmission attacks and impersonation attacks. If an eavesdropper uses any of these attacks, he can obtain the entire secret message, which means that not only part of the message has been leaked, but the entire message has been leaked. Furthermore, with impersonation attacks, legitimate parties cannot detect the presence of eavesdroppers. Therefore, it is necessary to modify the YZCSS protocol to improve its security.

YZCSS (Yan, Zhang, Chang, Sun, Sheng) protocol

In the YZCSS protocol [3], which has two sides Alice and Bob with their identificators ID_A and ID_B accordingly, where ID_A, $ID_B \in \{0, 1\}^N$. Alice sends a message $M \in \{0, 1\}^N$. Then Bob applies individual photons and Bell states, which are defined by the formulas:

$$|\phi\pm\rangle = \frac{1}{\sqrt{2}}\left(|00\rangle \, or \, |\phi^-\rangle\right),$$

$$|\varphi\pm\rangle = \frac{1}{\sqrt{2}}\left(|01\rangle \pm |10\rangle\right).$$

The stages of the protocol are as follows:

First stage. Alice and Bob have common identifiers ID_A and ID_B, use some QKD to exchange data. Alice begins the process of preparing two packets of data in two-qubit states S_M and S_A, according to M message and her own identification ID_A, each ordered set contains pairs of N qubits. For $1 \le i \le N$ let the i-th bit M (either ID_A or ID_B) will be M_i (either $ID_{A,i}$ or $ID_{B,i}$) and the i-th qubit S_M (або S_A) will be $S_{M,i}$ (or $S_{A,i}$). She prepares qubits using the following rule: (a) if M_i (or $ID_{A,i}$) = 0, so that $S_{M,i}$ (or $S_{A,i}$) = |01| or |10| with equal probability, (b) if M_i (or $ID_{A,i}$) = 1, so that $S_{M,i}$ (or $S_{A,i)}$ = $|\Phi^+|$ or $|\Phi^-|$ with equal probability.

Pairs of qubits of an ordered set S_A are called decoy states. Alice now inserts these state decoys into an ordered set S_M according to such a directive: (a) if $ID_{B,i} = 0$, then she inserts $S_{A,i}$ before $S_{M,i}$ and (b) if $ID_{B,i} = 1$, then she inserts $S_{A,i}$ after $S_{M,i}$.

Let the newly ordered set be S containing 2N pairs of qubits. Alice addresses S to Bob using a quantum communication channel. Consider an example:

Sample 1. Let's suppose that, M=10110, ID_A=01101 and ID_B=01001. Then

$$S_M = \left\{|\phi^+\rangle, |01\rangle, |\phi^+\rangle|\phi^-\rangle, |01\rangle\right\}, \; S_A = \left\{|10\rangle, |\phi^-\rangle, |\phi^-\rangle, |01\rangle, |\phi^+\rangle\right\} \text{ and}$$

$$S = \left\{|10\rangle, |\phi^+\rangle, |01\rangle, |\phi^-\rangle|\phi^-\rangle, |\phi^+\rangle, |01\rangle, |\phi^-\rangle, |01\rangle, |\phi^+\rangle\right\}.$$

Second stage. After Bob receives S, he knows the basis of the two photons corresponding to his ID. ID_B. Bob measures these decoy photons in the correct bases. If $ID_{A,i} = 0$, he chooses the basis Z × Z, where Z = {|0i, |1i}, so that Z × Z = {|00i, |01i, |10i, |11i}, and if $ID_{A,i} = 1$, then he chooses the Bell basis = {$|\Phi^+|$, $|\Phi^-|$, $|\Psi^+|$, $|\Psi^-|$} for measuring $S_{A,i}$. Bob also randomly measures pairs of qubits S_M in the basis Z × Z or in the Bell basis. After that, he notes the results of the measurements.

Third stage. Bob asks Alice to declare the starting states of the S_A qubit pairs for a security check. They compare the initial states and the decoy photon measurements and calculate the number of errors. If the number of errors is greater than the seventh predefined threshold, they stop the protocol, otherwise, they continue.

Fourth stage. Bob obtains all the bits of the secret message by calculating pairs of qubits S_M. The relationship between the measurement results and the bits of the secret message is shown in Table 1. Alice and Bob open seven parts of the received data to verify the integrity of the message.

The IZCSS protocol is robust against impersonation, interception, and retransmission attacks, man-in-the-middle attacks, etc. However, an eavesdropper can develop a strategy that allows him to perform an intercept and resend attack effectively.

Table 1.

Different cases of decoding the YZCSS protocol

Bits of Alice's secret message M_i	Encoded qubits $S_{m,i}$	Bases selected by Bob	Bob's measurement results	Decoded secret bits
0	$\|01\rangle$	bases $Z \times Z$	$\|01\rangle$	0
		Bell bases	$\|\psi^+\rangle or \|\psi^-\rangle$	0
	$\|10\rangle$	bases $Z \times Z$	$\|10\rangle$	0
		Bell bases	$\|\psi^+\rangle or \|\psi^-\rangle$	0
1	$\|\phi^+\rangle$	bases $Z \times Z$	$\|00\rangle or \|11\rangle$	1
		Bell bases	$\|\phi^+\rangle$	1
	$\|\phi^-\rangle$	bases $Z \times Z$	$\|00\rangle or \|11\rangle$	1
		Bell bases	$\|\phi^-\rangle$	1

Protocols using GHZ-like states

Recently, two very interesting DSQC protocols [4] based on particle reordering have been proposed. Yuan's protocol uses a four-qubit symmetric W state for communication, while Tsai's protocol uses dense coding of four-qubit cluster states. Current work is aimed at increasing the qubit efficiency of existing DSQC protocols and exploring the possibility of developing DSQC and QSDC protocols using GHZ-like and other quantum states.

GHZ-like states can be described generally as

$$\frac{\left(|\psi_i\rangle|0\rangle + |\psi_j\rangle|1\rangle\right)}{\sqrt{2}},$$

where $i, j \in \{0, 1, 2, 3\}$, $i \neq j$, also $|\psi_i|$ and $|\psi j|$ — Bell states, which are usually denoted as

$$|\psi_0\rangle = |\psi_{00}\rangle = |\psi^+\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}},\ |\psi_1\rangle = |\psi_{01}\rangle = |\psi^+\rangle = \frac{|01\rangle + |10\rangle}{\sqrt{2}},$$

$$|\psi_2\rangle = |\psi_{10}\rangle = |\psi^-\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}},\ |\psi_2\rangle = |\psi_{11}\rangle = |\psi^-\rangle = \frac{|01\rangle + |10\rangle}{\sqrt{2}}.$$

GHZ-like states are useful for controlled quantum teleportation. It is also possible to form an orthonormal basis set in 2^3 3-dimensional Hilbert spaces of 8 states that can be used for dense coding and DSQC. Thus, states are created as a useful resource for quantum information processing.

DSQC using GHZ-like states without using dense coding

Assume that Alice and Bob are two remote or spatially separated legitimate/authenticated communicators. Alice wants to give Bob a secret classic message. The proposed protocol [5] can be implemented using the following steps:

1. It can be assumed that Alice prepared n-copies of the GHZ-like state

$$|\lambda\rangle = \frac{|\phi^+ 0\rangle + |\phi^+ 1\rangle}{\sqrt{2}} = \frac{1}{2}\left(|010\rangle + |100\rangle + |001\rangle + |111\rangle\right).$$

Alice now prepares a sequence P of n-ordered triplets of entangled particles as $P = \{p_1, p_2 ....., p_n\}$, where index 1, 2, ..., n denotes the triplet order of particles $p_i = \{h_1, t_1, t_2\}$, which is in a state $|\lambda|$. Symbols h and t are used to denote the home photon (h) and the companion photon (t), respectively.

2. Alice encodes her secret message in the sequence P by applying one of four two-qubit unitary operations $\{U_{00} = X \otimes I,\ U_{01} = I \otimes I,\ U_{10} = I \otimes Z,\ U_{11} = I \otimes iY\}$ into particles (h_1, t_1) of each triplet. Unitary operations $\{U_{00}, U_{01}, U_{10}, U_{11}\}$ encodes a secret message {00, 01, 10, 11} respectively. Here [5]

$$I = |0\rangle\langle 0| + |1\rangle\langle 1|,$$

$$X = \sigma_x = |0\rangle\langle 1| + |1\rangle\langle 0|,$$

$$iY = i\sigma_y = |0\rangle\langle 1| - |1\rangle\langle 0|,$$

$$Z = \sigma_z = |0\rangle\langle 0| - |1\rangle\langle 1|.$$

These operations U_{ij} (i, j∈{0, 1}) will transform a GHZ-like state |λ| into another GHZ-like state $|\lambda_{ij}|$, where

$$|\lambda_{00}\rangle = U_{00}|\lambda\rangle = \frac{1}{2} X \otimes I\left(|010\rangle + |100\rangle + |001\rangle + |111\rangle\right) =$$

$$\frac{1}{2}\left(|010\rangle + |000\rangle + |101\rangle + |011\rangle\right) = \frac{|0\psi^+\rangle + |1\phi^+\rangle}{\sqrt{2}}.$$

$$|\lambda_{01}\rangle = U_{01}|\lambda\rangle = \frac{1}{2} I \otimes I\left(|010\rangle + |100\rangle + |001\rangle + |111\rangle\right) =$$

$$\frac{1}{2}\left(|010\rangle + |100\rangle + |001\rangle + |111\rangle\right) = \frac{|0\phi^+\rangle + |1\psi^+\rangle}{\sqrt{2}}.$$

$$|\lambda_{10}\rangle = U_{10}|\lambda\rangle = \frac{1}{2} I \otimes Z\left(|010\rangle + |100\rangle + |001\rangle + |111\rangle\right) =$$

$$\frac{1}{2}\left(-|010\rangle + |100\rangle + |001\rangle - |111\rangle\right) = \frac{|0\phi^-\rangle + |1\psi^-\rangle}{\sqrt{2}}.$$

$$|\lambda_{11}\rangle = U_{11}|\lambda\rangle = \frac{1}{2} I \otimes iY\left(|010\rangle + |100\rangle + |001\rangle + |111\rangle\right) =$$

$$\frac{1}{2}\left(-|000\rangle + |110\rangle + |011\rangle - |101\rangle\right) = \frac{\left(|0\psi^-\rangle + |1\phi^-\rangle\right)}{\sqrt{2}}.$$

3. Alice stores the home photon (h1) of each triplet and prepares an ordered sequence, $P_A = [p_1(h_1), p_2(h_1), ..., p_n(h_1)]$. Similarly, it uses all traveling photons to prepare an ordered sequence $P_B = [p_1(t_1, t_2), p_2(t_1, t_2), ..., p_n(t_1, t_2)]$.

4. Alice disrupts the order of a pair of traveling photons in P_B and creates a new sequence $P'_B = [p'_1(t_1, t_2), p'_2(t_1, t_2), ..., p'_n(t_1, t_2)]$. The actual order is known only to Alice.

5. To prevent eavesdropping, Alice prepares m = 2n decoy photons. Decoy photons are randomly prepared in one of four states {|0|, |1|, |+|, |−|}, where |+|=|0|+|1|/√2 and |−| =|0|−|1/|√2, that is, the state of decoy photons $\bigotimes_{j=1}^{m}|P_j|$, $|P_j| \in \{|0|, |1|, |+|, |-|\}$, (j = 1, 2, ..., m). Alice then randomly inserts these decoy photons into the sequence P'_B and creates a new sequence P'_{B+m}, which she hands to Bob, P_A stays with Alice.

6. After confirming that Bob has received the entire sequence P'_{B+m}, Alice announces the positions of decoy photons. Bob measures the corresponding particles in sequence P'_{B+m} using base X or Z randomly, where X = {|+|, |−|} and Z = {|0|, |1|}. After the measurement, Bob publicly announces his result and the basis used for the measurement. Alice now has to reject 50% of the times Bob chose the wrong basis. From the remaining results, Alice can calculate the error rate and check whether it exceeds a predefined threshold or not. If it exceeds the threshold, Alice and Bob stop this communication and repeat the procedure from the beginning. Otherwise, they proceed to the next step.

7. Knowing the position of the decoy photons, Bob already had the sequence P'_B, Alice reveals the actual order of the sequence, and Bob uses this information to transform the reordered sequence P'_B to the original sequence P_B. Therefore, Alice needs to exchange 2n classical bits.

8. Alice measures photons on a computational basis (Z basis) and announces the result. Bob measures the received qubits in the Bell base. Knowing the results of Alice's measurements and his measurements, Bob can easily decode the encoded information. For clarity, in Table 1, we have provided the relationship between measurement results and secret messages.

Table 2.

Relationship between measurement results and secret message in DSQC using GHZ-like states without full use of dense coding

Alice's measurement result	Bob's measurement result	Decoded secret
0	ϕ^+	01
	ϕ^-	10
	ψ^+	00
	ψ^-	11
1	ϕ^+	00
	ϕ^-	11
	ψ^+	01
	ψ^-	10

A similar DSQC protocol based on particle rearrangement was recently proposed by Yuan and the others [6]. In their work, they used a four-qubit symmetric state W to securely transmit 2 bits of classical information. They compared their protocol with the previous DSQC protocol proposed by Cao and Song. Their protocol also uses 4-qubit W-states for DSQC, but each of these W-states can only be used to transmit one bit of classical information. With this in mind, Yuan et al claim that their protocol has high throughput because each W state can transport two bits of information that is secret[7]. The advantage is that the encoding is performed by performing a single operation on two qubits (photons), but only one of them is stored as the master photon. The same conclusion requires the GHZ state and any other significantly densely encoded tripartite state.

Summarizing, it is possible to demonstrate weak positions for each of the mentioned protocols, as well as to show which cyber-attacks they are resistant to.

Table 3.

Presentation of the resistance of protocols to different types of cyber-attacks, where "+" means resistance, and "-" indicates vulnerability to such attacks

QSDC Protocols	The main types of cyber attacks								
	C	NC	MiM	DoS	TC	FS	PBS	PNS	TH
Ping-Pong, DLL, PP^{GV}	+	–	–	–	+	+	+	+	–
Entangled qubits in groups	+	+	–	–	+	+	+	+	–
With groups of confused qudites	+	+	–	–	+	+	+	+	–
With single qubits	+	+	–	–	+	+	+	+	–
YZCSS	+	+	–	–	+	+	+	+	–
using GHZ-like states	+	+	–	–	+	+	+	+	–
using GHZ-like states without using dense coding	+	+	–	–	+	+	+	+	–

Conclusions

Thus, this paper analyzes modern protocols of quantum cryptography (identifies their advantages and disadvantages), and their existing classifications. Based on partial generalizations of theoretical provisions and practical achievements in the field of quantum cryptography, a generalized classification was developed. By comparing various factors of the protocols, and their resistance to certain cyberattacks, we have the opportunity to identify several problems in this field and expand the possibilities for choosing appropriate methods for building modern quantum information protection systems.

References

1. Zhmurko T., Kinzeryavyy V., Yubuzova Kh., Stojanovic A.: Generalized classification of modern quantum cryptography and communication methods. Ukrainian Scientific Journal of Information Security, 2015, vol. 22, issue 3, p. 287-293.
2. Banerjee A., Pathak A.: Efficient protocols for deterministic secure quantum communication using GHZ-like states. Quantum Physics, 2018.
3. Yan L., Zhang S., Chang Y., Sun Z., Sheng Z.: Quantum secure direct communication protocol with mutual authentication based on single photons and Bell states. Computers, Materials & Continua, 63(3):1297–1307, 2020.
4. Banu N., Ghosal P, Panigrahi P. K.: "Quantum information splitting of an unknown two qubit state by using two three qubit GHZ like states," 2014 International Conference on Electronics and Communication Systems (ICECS), 2014, pp. 1-4, doi: 10.1109/ECS.2014.6892773.
5. Guo W., Hou X.: An Efficient Controlled Quantum Secure Direct Communication Protocol via GHZ-like States. 2019 IEEE 5th International Conference on Computer and Communications (ICCC), 2019, pp. 821-825, doi: 10.1109/ICCC47050.2019.9064457.
6. Shukla Ch., Banerjee A., Pathak A.: Improved Protocols of Secure Quantum Communication Using W States. International Journal of Theoretical Physics, 2018, vol. 52, pp. 1914–1924.
7. Yuan, H., Song, J., Zhou, J. et al. High-capacity Deterministic Secure Four-qubit W State Protocol for Quantum Communication Based on Order Rearrangement of Particle Pairs. Int J Theor Phys 50, 2403–2409 (2011). https://doi.org/10.1007/s10773-011-0729-7

Tetiana Okhrimenko **Тетяна Охріменко**	PhD, Senior Research Fellow, Research Laboratory of Cyber Threats Counteraction in Aviation, National Aviation University, Kyiv, Ukraine, e-mail: t.okhrimenko@npp.nau.edua.ua https://orcid.org/0000-0001-9036-6556	к.т.н., страший науковий співробітник науково-дослідної лабораторії протидії кіберзагрозам в авіаційній галузі, Національний авіаційний університет, Київ, Україна
Serhii Dorozhynskyi **Сергій Дорожинський**	PhD-student, assistant of the department of telecommunications and radioelectronic systems, young scientist of scientific research department, National Aviation University, Kyiv, Ukraine, **e**-mail: dorozhun1706@gmail.com. https://orcid.org/0000-0002-5395-6423	PhD-аспірант, асистент кафедри телекомунікацій та радіоелектронних систем, молодший науковий співробітник науково-дослідної частини, Національний авіаційний університет, Київ, Україна
Bohdan Horbakha **Богдан Горбаха**	Graduate student, Laboratory assistant of scientific research department, National Aviation University, Kyiv, Ukraine e-mail: 4591078@stud.nau.edu.ua, https://orcid.org/0000-0003-0713-4426	студент магістратури, Лаборант науково-дослідної частини, Національний Авіаційний Університет, Київ, Україна

UDC 004.89: 004.3
https://doi.org/10.31891/csit-2023-1-9

Olga PAVLOVA, Andriy BASHTA, Mykola KOVTONIUK
Khmelnytskyi National University

AUGMENTED REALITY BASED INFORMATION TECHNOLOGY FOR OBJECTS 3D MODELS VISUALIZATION

At the current stage of IT industry development, augmented reality is of interest both from the side of science and from the business side, since it is an advanced and newest tool for introducing a new immersive user experience. Today there are plenty ready-to-use applications that use AR for business, educational, medical and other purposes. Augmented Reality is currently one of the most popular upcoming technologies most commonly known for its use within games and advertising. By combining three-dimensional modelling with augmented reality, it could be possible to obtain new user friendly applications for the representing 3D models of objects in real time and in real size. The topic of research in the field of augmented reality is currently relevant both for science and for the business industry.

The paper proposes a multifunctional information system for three-dimensional models visualization in augmented reality, which is implemented in the form of a cross-platform mobile application. The proposed information system uses a device camera as a mean of object visualization and provides quick reproduction of the selected from the application`s database model in augmented reality in real size and in real time.

The developed application works quite well, has a user friendly and intuitive interface and allows user to add own models, that makes this tool multipurpose.Test 3D models have been created for conducting experiments for verification the proposed information system operation.

The further efforts of the authors will be directed to improving the existing algorithms for extending the current functionality of the proposed tool for 3D objects models visualization in augmented reality and application of the developed tool for real-life needs, such as digitization and visualization of museum exhibits and archaeological artifacts of Khmelnytskyi region.

Keywords: information system, Augmented Reality (AR), 3D model, visualization,cross-platform mobile application.

Ольга ПАВЛОВА, Андрій БАШТА, Микола КОВТОНЮК
Хмельницький національний університет

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ДЛЯ ВІЗУАЛІЗАЦІЇ 3D-МОДЕЛЕЙ ОБ'ЄКТІВ У ДОПОВНЕНІЙ РЕАЛЬНОСТІ

На сучасному етапі розвитку інформаційних технологій доповнена реальність становить інтерес як з боку науки, так і з боку бізнесу, адже є передовим та новітнім інструментом для впровадження нового досвіду користувача. Наразі вже є багато розроблених застосунків із використанням доповненої реальності для комерційної, освітньої, медичної та інших сфер застосування. Доповнена реальність сьогодні є однією з перспективних технологій, яка відома своїм застосуванням у рекламі та ігровій індустрії. Через поєднання 3D моделювання та доповненої реальності стало можливо створити новий зручний застосунок для відтворення тривимірних моделей у доданій реальності у режимі реального часу та в реальному розмірі. Тема досліджень у сфері доповненої реальності наразі є актуальною як для науки, так і для бізнес-індустрії.

У роботі запропоновано багатофункціональну інформаційну систему для візуалізації тривимірних моделей у доповненій реальності, яка реалізована у вигляді кросплатформного мобільного додатку. Запропонована інформаційна система використовує камеру пристрою як засіб візуалізації об'єкта та забезпечує швидке відтворення обраної з бази даних додатку моделі в доповненій реальності в реальному розмірі та в режимі реального часу.

Розроблений додаток працює досить добре, має зручний та інтуїтивно зрозумілий інтерфейс і дозволяє користувачеві додавати власні моделі, що робить цей інструмент багатоцільовим. Також було створено тестові 3D моделі для проведення експериментім для перевірки роботи пропонованої інформаційної системи.

Подальші зусилля авторів будуть спрямовані на вдосконалення існуючих алгоритмів для розширення поточної функціональності запропонованого інструменту для візуалізації 3D моделей об'єктів у доповненій реальності та застосування розробленого інструменту для реальних потреб, таких як оцифрування та візуалізація музейних експонатів та археологічних артефактів Хмельницької області.

Ключові слова: інформаційна система, доповнена реальність (AR), 3D-модель, кросплатформний мобільний застосунок

Introduction

Currently, augmented reality technology is gaining more and more frequent use and is applied to more areas of human life. We can see mobile applications, commercial and educational simulators, mobile games and even the advertisement using augmented reality. AR technology has huge commercial potential in a variety of industries, from opening new marketing channels to improving employee training processes. According to the information of the International statistical company Statista [1], in recent years there has been a significant increase in the number of software products using augmented reality. According to forecasts [2], by 2024 there will be around 2.4 billion augmented reality applications users in the world. This technology is expected to reach $70-$75 billion in revenue by 2024. The diagram (Fig. 1) shows the growth dynamics of the number of software products based on augmented reality, both for commercial use and user applications. The graph also shows positive dynamics from an economic point of view, since the development and use of AR-based applications means the growth of global profits.

Therefore, the topic of research in the field of augmented reality is currently relevant both for science and for the business industry.

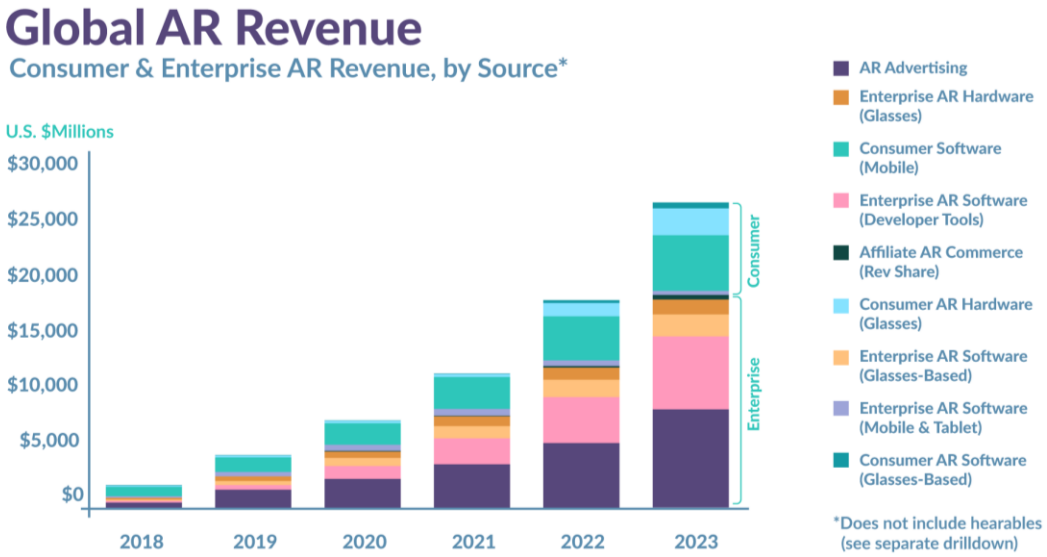


Fig.1.The prospects for AR industry development in the world [2]

Domain analysis and literature review

According to market research [3-4], the industries that the most actively leverage AR technology are E-commerce and advertising (IKEA Place and WannaKicks – the applications that provide virtual products fitting and try on using AR), E-learning (Civilizations - an application launched by the BBC for British Museum; provides visualization of ancient historical artifactsin AR) gaming industry (AR-games ZombiesRun and PokemonGO), medicine and healthcare (SentiAR – the platform features a 3D visualization of a patient's heart anatomy to assist with diagnostics and surgerie), social networking (Instagram ans Snapchat AR-filters). The integration of augmented reality technology in various areas of modern society is presented in Fig.2

Fig.2.The examples of augmented reality technology integration in various areas of modern society

During the research the analysis of the existing trends and directions of augmented reality application for different areas of life has been conducted. The results of the analysis are presented in Table 1.

The analysis of scientific publications for 2021-2022 and already existing solutions presented on the world market showed that augmented reality is most often used for educational simulators, in immersive technologies, in the field of advertising, e-commerce, for the development of the latest VR/MR user interface. The results of the recent scientific publications topics using augmented reality are shown on a diagram in Fig. 3.

In Ukraine, augmented reality domain is also of interest to scientists. In recent years, the number of publications on the application of augmented reality for the field of education has increased significantly [6-12]. Thus, the paper [6] presentst an analysis of the current state and prospects for the development of augmented reality in

Ukraine in business and education. The authors of [7] propose the application of augmented reality for educational purposes. In [8] using augmented reality-based technologies in professional training of future teachers of Ukrainian language and literature is proposed. The authors of [9] propose augmented reality-based approach for immersive training for some specific professions. In [10] augmented reality is proposed to be used in university education of future IT specialists. The source [11] considers application of augmented reality as an interactive form of pre-school and primary school teaching. The authors of [12] propose using AR technology for visualization of atoms and moleculas structures at Chemistry lessons. The paper [13] is devoted to using of augmented reality for navigation and paving routes in real time. However *none of the considered works is devoted to application of AR technology for three-dimensional objects visualization.*

Table 1

Trends and directions of augmented reality application for different areas of life

Technology or Trend of AR application	Application or Device	Description
AR-based virtual companions	Hybri virtual companion [4]	Such an app will allow users to create an AR-based AI companion which resembles real-life humans.
Leverage AR glasses for different use cases	Google company	Giants like Apple are already working on bringing unique AR experience through glasses slated to release in late 2025.
Cash in on the metaverse	Microsoft Azure	Microsoft is one of the technology giants looking to dominate the AR experience. Their concept of merging cloud computing with AR/VR excellence makes it an interesting case study.
Mobile Augmented Reality	Instagram Snapchat	One of the finest examples of mobile AR is the image filter users can use in social media apps like Snapchat and Instagram. Snapchat is preparing to introduce NFT(Non-fungible tokens) as filters in their mobile applications.
Remote collaborations with AR	-	One fine example is cryogenic operations in oil refineries, where temperature maintenance is critical. While on-site employees come across issues that need expert assistance, AR can help with necessary response assistance.
Immersive AR gaming applications	AR games like Pokemon GO and Egg.	One of the critical AR trends is the use of Augmented Reality in creating interactive entertainment for users. In addition, Inc and Harry Potter the wizards unite have provided enhanced gaming experience on smartphones.
AR-based marketing & advertising campaigns	Loreal Burberry	Youtube has helped businesses with a “beauty try-on” feature that allows cosmetic brands to let users experience the product.Following the AR market trend, brands like Loreal has already created unique and immersive shopping experience in their apps. Not just as a marketing approach, AR is an excellent way to advertise your products. For example, Burberry allows its customers to create a custom handbag AR model, which they can experience through the web app. As a result, brands can leverage AR technology and create a customized experience for their customers.
Supply chain efficiency improvementwith AR	Walmart	For example, Walmart converted four of its physical retail stores into the testing environment for an AR-based inventory app. The idea was to reduce the time needed for bringing products from backroom inventory to the sales floor.

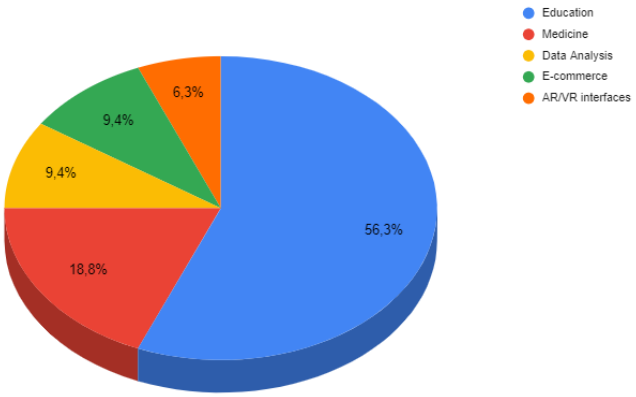


Fig.3. The results of the recent scientific publications topics using augmented reality

Thus, taking into account the relevance and importance of this task, the purpose of this work is to develop an information system for visualization of three-dimensional objects based on augmented reality in the form of a cross-platform mobile application and conduct experiments for visualization of 3D models of objects using the proposed information technology.

Augmented reality based information technology for objects 3D models visualization

In our previous works, a method and algorithm of information technology work for visualization of three-dimensional objects based on AR were proposed. For developing the information system for objects 3D models visualization, we will be using smartphone camera as a tool for representing the objects in augmented reality. To present a 3D model in a real world space, it is necessary to consider that we need to work with three-dimensional space.

The smartphone must also be located at the intersection of the X, Y and Z axes in the 3D Cartesian coordinate system as shown in Fig. 4

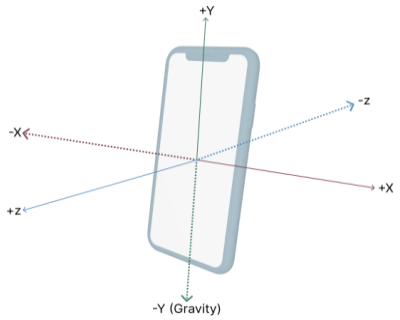


Fig.4. Location of the user's smartphone in the three-dimensional Cartesian coordinate system

The principle of operation of the proposed information system is that the user enters the mobile application. Since the application is cross-platform, it provides visualization of models from both Android and iOS devices. Next, the user can either upload his own model to the system database or choose one of the proposed ready-to-visualize models contained in the system database. Once selected, the model is available to the user for preview with the option to display in augmented reality in real time. The user can adjust the size and position of the model in the field of view of the smartphone camera and take a photo or video of the model in the environment. The principle of the proposed information system operation is presented in Fig.5.

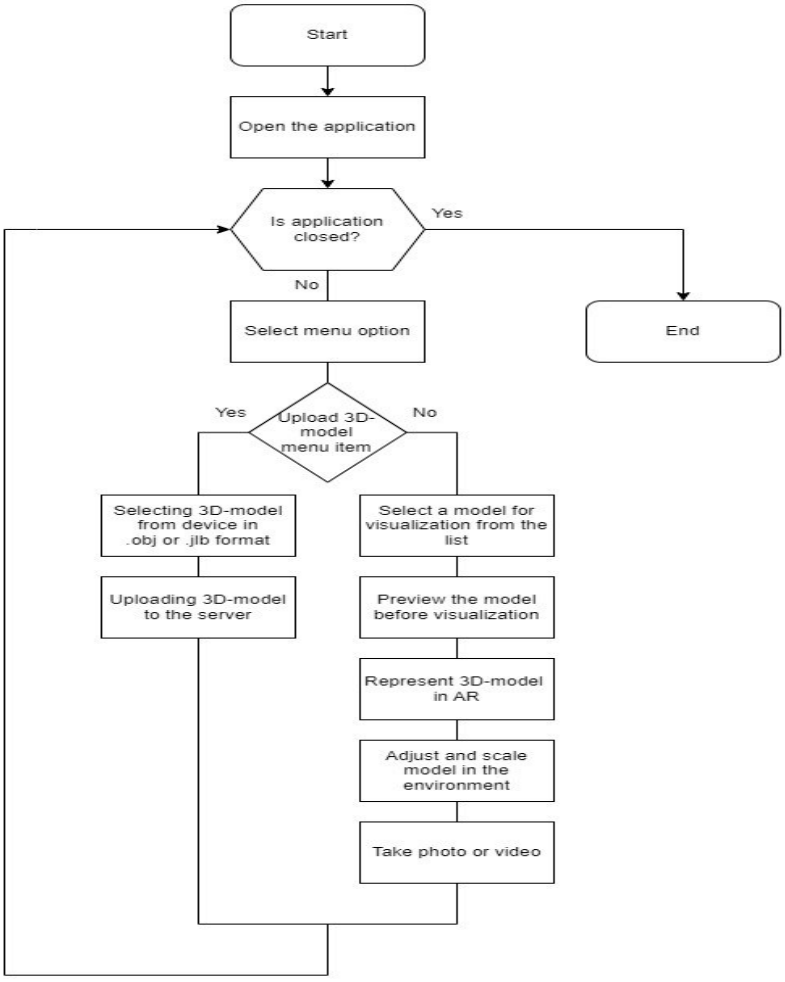
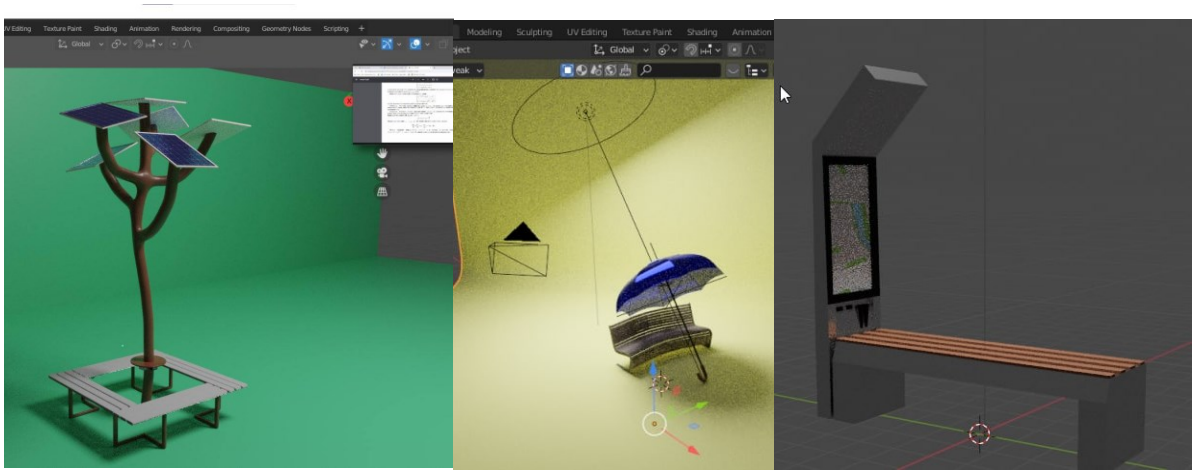


Fig.5. L The principle of operation of the information system for 3D models visualization in AR

Experiments and Discussion

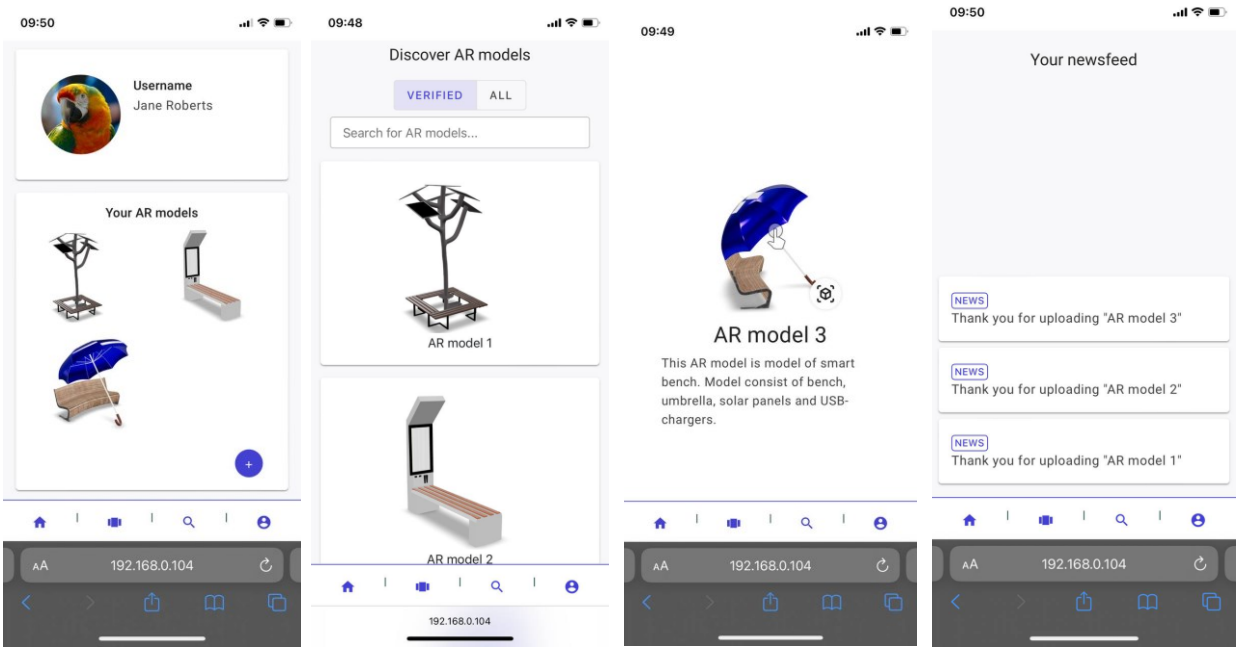
To be able to conduct the experiments, we need to create the 3D models for testing. We chose the Botanical Garden of Khmelnytskyi National University as an environment for displaying the models and decided to create the models of the outdoor smart art objects than can be represented in AR in real size and in real time.We chose Blender as an environment for 3D models creating. The models are presented in Fig.6.



a) b) c)

Fig. 6. 3D Models of art structures for Information Technology for 3D objects models visualization testing Smart Solar Pannels Tree condtruction; b) The bench with Smart Solar Pannels Umbrella; c) Smart Bench with Solar Pannel based Information Screen)

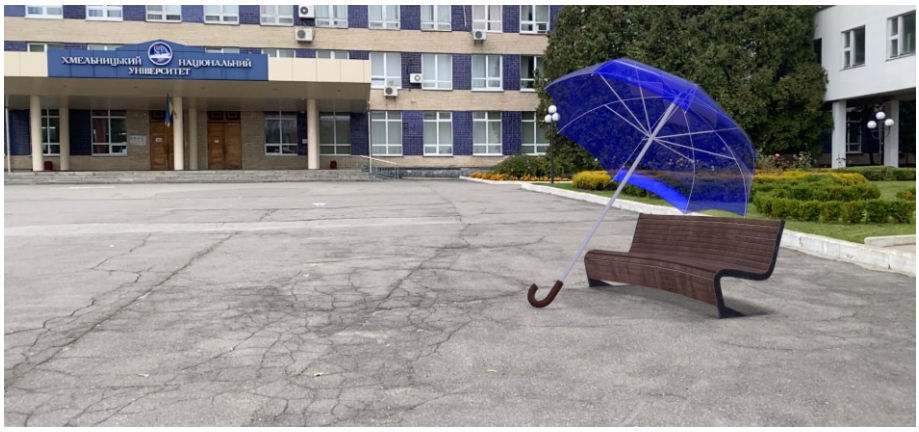
For conducting the experiments the proposed Information Technology for 3D objects models visualization has been developed in the form of a cross-platform mobile application and installed on Andriod OS-driven mobile device. The interface screens of the developed information system are presented in Fig.7.



a) b) c) d)

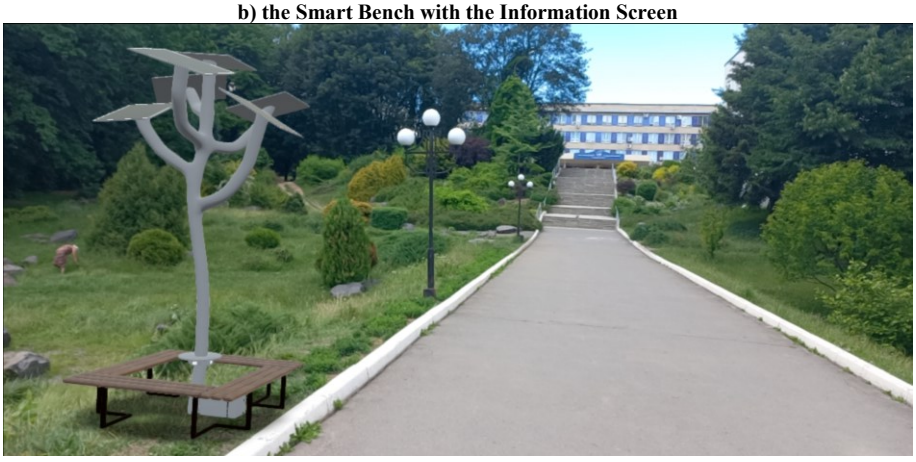
Fig. 7. Interface screens of Information Technology for 3D objects models visualization User Profile Interface; b) Set of the available AR models; c) Preview of the selected model with the description before the visualization; d) The process of models uploading)

The results of the experiments are presented as a set of photos taken with the smartphone camera (Fig 8). The objects on photos are placed in the environment (Khmelnytskyi National University campus and Botanical garden) in real size.



a) the Bench and the Smart Umbrella

b) the Smart Bench with the Information Screen



c) the Solar Tree Art-object

Fig. 8. The experiment results on 3D objects models visualization in AR

Conclusions

Augmented reality has always been an object of interest as scientists as business industry representatives. The prospect of this technology application for various areas of industry and social life is proved by the numerous research works as well as the increasing number of commercial program products.

During the study the conducted literature analysis and the analysis of already existing AR-based mobile applications provided the conclusions that currently there are no ready-to-use solutions that provide reproducing a three-dimensional model of an object in augmented reality, so this is currently an urgent task from both a scientific and a practical point of view.

Therefore the information system for 3D objects models visualization in augmented reality was developed in the form of cross-platform mobile application. The proposed information system uses a device camera as a mean of object visualization and provides quick reproduction of the selected from the application`s database model in augmented reality in real size and in real time.

The developed application works quite well, has a user friendly and intuitive interface and allows user to add own models, that makes this tool multipurpose.The further efforts of the authors will be directed to improving the existing algorithms for extending the current functionality of the proposed tool for 3D objects models visualization in

augmented reality and application of the developed tool for real-life needs, such as digitization and visualization of museum exhibits and archaeological artifacts of Khmelnytskyi region.

References

1. Number of mobile augmented reality (AR) active user devices worldwide from 2019 to 2024. URL: https://www.statista.com/statistics/1098630/global-mobile-augmented-reality-ar-users/ (accessed February 13, 2022)
2. Augmented Reality in Business: Benefits. URL: https://redwerk.com/blog/augmented-reality-in-business/ (accessed February 13, 2022)
3. 10 Augmented Reality Trends to Redefine Business Growth in 2023. URL: https://www.intelivita.com/blog/augmented-reality-trends/ (accessed February 14, 2022)
4. 12 Augmented Reality Trends of 2023: New Milestones in Immersive Technology. URL: https://mobidev.biz/blog/augmented-reality-trends-future-ar-technologies (accessed February 14, 2022)
5. 'Hybri' Can Create a Virtual Companion Based on Real People URL:https://www.macobserver.com/link/hybri-virtual-companions/ (accessed February 3, 2022)
6. Mitii I., & Soloviov V. (2018). Augmented Reality: Ukrainian Present Business and Future Education. *Educational Dimension*, *51*, pp. 290-296.
7. Iatsyshyn A., et al. Application of augmented reality technologies for education projects preparation. *CTE Workshop Proceedings.* Vol. 7. 2020. pp. 134-160
8. Petrovych O., et al. The usage of augmented reality technologies in professional training of future teachers of Ukrainian language and literature. CEUR Workshop Proceedings, 2021.
9. Iatsyshyn A., et al. Application of augmented reality technologies for preparation of specialists of new technological era. (2020).
10. Babkin V., et al. Using augmented reality in university education for future IT specialists: educational process and student research work. CEUR Workshop Proceedings, 2021.
11. Palamar S., et al. Formation of readiness of future teachers to use augmented reality in the educational process of preschool and primary education. CEUR Workshop Proceedings, 2021.
12. Nechypurenko P., Stoliarenko V., Starova T., Selivanova T., Markova O., Modlo Y., & Shmeltser E. (2020). Development and implementation of educational resources in chemistry with elements of augmented reality.
13. Pavlova O., Bashta A., Kravchuk S., Hnatchuk Y., Bouhissi H.E. Augmented Reality Based Technology and Scenarios for Route Planning and Visualization. CEUR Workshop Proceedings, 2022, 3156, pp. 613–623

Ольга Павлова **Olga Pavlova**	PhD, Senior Lecturer of Computer Engineering & Information Systems Department, Khmelnytskyi National University, Khmelnytskyi, Ukraine, e-mail: pavlovao@khmnu.edu.ua https://orcid.org/0000-0001-7019-0354	доктор філософії, старший викладач кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет, Хмельницький, Україна
Андрій Башта **Andriy Bashta**	PhD student of Computer Engineering & Information Systems Department, Khmelnytskyi National University, Khmelnytskyi, Ukraine, e-mail: andreybashta@gmail.com https://orcid.org/0000-0002-0775-1347	аспірант кафедри комп'ютерної інженерії та інформаційних ситем, Хмельницький національний університет, Хмельницький, Україна
Микола Ковтонюк **Mykola Kovtoniuk**	Master student of Computer Engineering & Information Systems Department, Khmelnytskyi National University, Khmelnytskyi, Ukraine, e-mail: mykola.kovtonuk@gmail.com https://orcid.org/0000-0001-7272-767X	студент магістратури за спеціальністю «Комп'ютерна інженерія», Хмельницький національний університет, Хмельницький, Україна

UDC 004.85:330.4
https://doi.org/10.31891/csit-2023-1-10

Vasyl PRYIMAK, Bohdan BARTKIV, Olga HOLUBNYK
Ivan Franko National University of Lviv

FORECASTING THE EXCHANGE RATE OF THE UKRAINIAN HRYVNIA USING MACHINE LEARNING METHODS

This article describes the concept of currency exchange rate and the typology of various factors that influence it. A multifactor regression model was constructed to investigate the influence of factors on the exchange rate of the Ukrainian hryvnia and to forecast the dynamics of this rate based on the studied factors using Data Science technologies.

The purpose of this work is to study the peculiarities of the formation of the exchange rate of the Ukrainian hryvnia, the characteristics of the influence of various external factors on this rate, and the creation of an effective forecasting model of the Ukrainian national currency rate, based on a certain number of fundamental financial and economic factors that influence this rate.

Macroeconomic indicators that theoretically have an impact on the dynamics of the currency exchange rate were chosen to build the model. Data on the exchange rate of the Ukrainian hryvnia to the US dollar and economic indicators for selected factors were collected from 2010 to September 2022. During the implementation of the task, the collected data was processed, brought into a uniform form, and normalized. Machine learning methods were used for regression modeling, specifically the XGBoost gradient boosting method.

As a result, a retrospective forecast of the Ukrainian hryvnia exchange rate was obtained, based on factor variables, and an estimate of the impact of each selected feature on the currency exchange rate was calculated. The scientific novelty of this work lies in the application of modern machine learning methods and technologies for the analysis, modeling, and forecasting of the exchange rate of the Ukrainian national currency.

The practical significance of this article lies in the possibility of using the proposed approaches to forecasting the exchange rate of the Ukrainian hryvnia with the use of machine learning methods by all interested parties, including financial institutions of Ukraine, to achieve stability of the national currency, which in turn will affect the development of the national economy as a whole and the welfare of the population of the country.

Keywords: exchange rate, gradient boosting, regression analysis, machine learning, forecasting, Ukrainian hryvnia, Data Science.

Василь ПРИЙМАК, Богдан БАРТКІВ, Ольга ГОЛУБНИК
Львівський національний університет імені Івана Франка

ПРОГНОЗУВАННЯ ВАЛЮТНОГО КУРСУ УКРАЇНСЬКОЇ ГРИВНІ З ВИКОРИСТАННЯМ МЕТОДІВ МАШИННОГО НАВЧАННЯ

У даній статті описано поняття валютного курсу та типологію різноманітних чинників впливу на нього. Побудовано багатофакторну регресійну модель для дослідження впливу факторів на курс української гривні та спрогнозовано динаміку цього курсу на основі досліджуваних факторів.

Метою даної роботи є дослідження особливостей формування валютного курсу української гривні, характеристика впливу різноманітних зовнішніх факторів на цей курс та створення за допомогою Data Science технологій ефективної моделі прогнозування курсу української національної валюти, яка ґрунтується на певній кількості фундаментальних фінансово-економічних факторів впливу на цей курс.

Для побудови моделі обрано макроекономічні показники, які теоретично мають вплив на динаміку валютного курсу. Для статистичного аналізу та подальшого моделювання зібрано дані про валютний курс української гривні до долара США та економічні показники для обраних факторних ознак за період з 2010 по вересень 2022 рр. В ході реалізації поставленого завдання проведено обробку, зведення до єдиної форми та нормалізацію зібраних даних. Для безпосереднього моделювання використано методи машинного навчання для задачі регресії, а саме метод градієнтного бустингу (XGBoost). В результаті отримано ретроспективний прогноз курсу української гривні, базований на факторних змінних і розраховано оцінку впливу кожної вибраної ознаки на валютний курс.

Наукова новизна даної роботи полягає у застосуванні сучасних методів та технологій машинного навчання для аналізу, моделювання та прогнозування курсу української національної валюти.

Практична значимість цієї статті полягає у можливості використання запропонованих у ній підходів до прогнозування валютного курсу української гривні з застосуванням методів машинного навчання всіма зацікавленими сторонами, зокрема фінансовими установами України, задля досягнення стабільності національної грошової одиниці, що у свою чергу вплине на розвиток національної економіки у цілому та добробут населення держави.

Ключові слова: валютний курс, градієнтний бустинг, регресійний аналіз, машинне навчання, прогнозування, українська гривня, Data Science.

Introduction

The exchange rate, despite being a measure of the value of the national currency expressed in monetary units of other countries, is also an indicator of the domestic economic situation, reflecting the main trends in the development of the national economy and influencing the redistribution of national income between countries.

Significant volatility in the national currency exchange rate can have negative consequences for export potential, foreign trade, and the economy as a whole. Therefore, the key task of state regulation of the monetary system in the conditions of a market economy is to study the factors that affect the exchange rate and react promptly to the main trends in the economy to ensure its stability.

Macroeconomic forecasting of economic indicators and processes, including forecasting of exchange rates, is a complex task. There are currently no macroeconomic models that would have the functionality to make reliable macroeconomic forecasts of economic development and the exchange rate of the national currency. Therefore, the development of such models is always an important and relevant problem.

The article examines the peculiarities of the formation of the exchange rate and analyzes the impact of various external factors on the exchange rate of the Ukrainian hryvnia. Using machine learning methods, a regression model, namely the gradient boosting model (XgBoost), was built to study the impact of the given factors on the exchange rate of the Ukrainian hryvnia. Modern Data Science technologies for data analysis and solving tasks of economic-mathematical modeling and socio-economic forecasting are considered.

Related works

This work analyzes the works of Ukrainian scientists on the factors influencing the exchange rate in general, and the exchange rate of the Ukrainian hryvnia in particular. The authors of scientific papers [1-3] define the concept of exchange rate and classify the factors that influence its formation. In the work [4], the author presents theoretical methods for predicting currency exchange rates. The basic principles of macroeconomic modeling and forecasting of the exchange rate in Ukraine are studied in the monograph [5]. In scientific papers [6-9], the authors investigate the factors influencing the formation of the exchange rate in Ukraine and analyze the degree of influence of various indicators on the exchange rate of the Ukrainian hryvnia as an integral part of the national economy.

The author of the scientific article [10] considers the main directions of using Data Science algorithms in central banks, including predicting macroeconomic and financial variables. The theoretical foundations of Data Science technologies and their application in economic-mathematical modeling are described in works [11-13]. The characteristics of machine learning algorithms for solving regression, classification, and forecasting tasks are presented in articles [14-15]. The authors of scientific papers [16-18] describe the principle of the gradient boosting modeling algorithm, its specificity, and the features of its application.

However, modeling and forecasting of the exchange rate are always relevant tasks, as new data appears every day, and trends in the economy change. Therefore, the purpose of this work is to study the peculiarities of the formation of the exchange rate of the Ukrainian hryvnia, the characteristics of the influence of various external factors on this rate, and to create an effective model for predicting the exchange rate of the Ukrainian national currency using Data Science technologies, based on fundamental financial and economic factors that affect it.

Presenting main material

The exchange rate is determined by the market interaction of demand and supply under conditions of perfect competition, reflecting a complex set of factors that directly and indirectly affect the exchange rate of both the national economy and international economic relations.

The multifactorial nature of the exchange rate reflects its connection with other economic categories, such as value, price, money, interest rates, and more. In modern conditions, the exchange rate is formed under the influence of demand and supply in the foreign exchange market, but along with the state of the balance of payments, its size is influenced by a large number of other factors, such as the level and dynamics of inflation, the amount of money in circulation, interest rates, GDP volumes, and growth rates, the level of development of the financial market, political and psychological factors, and much more. As a result, the formation of the exchange rate at the present stage is considered a multifactorial process. Modern researchers of the process of exchange rate formation group numerous exchange rate factors according to certain characteristics. In particular, factors are divided into three groups: fundamental, technical, and force majeure. Fundamental factors are key macroeconomic indicators of the state of the national economy that affect foreign exchange market participants and the exchange rate level.

Figure 1 provides a more detailed demonstration of the variety of factors that influence the exchange rate. Such a distribution is quite conditional because some components cannot be unequivocally attributed to a certain group, as most of them are interrelated. However, this can facilitate the perception of the entire complex of components forming the exchange rate of the currency.

Among the social and political factors, one can distinguish the political situation in the country, the level of trust of the population in the banking system, the presence of a "black market," the level of financial literacy of the population, and others. Additionally, to some extent, the psychological factor that shapes public attitudes based on forecasts of currency exchange rates spread by rumors, forecasts, and speculation in the media, which generate excitement in the currency market, also influences the exchange rate.

In the current conditions of an unstable economic situation, such a factor as speculative capital flows deserves special attention. This factor can affect the dynamics of the exchange rate if the central bank tries to keep it at a certain level against the action of market forces. If the exchange rate of a certain national currency tends to decrease, firms and banks try to sell it in advance in exchange for more stable currencies, counting on conducting a reverse operation at a lower rate after a certain period. The difference, therefore, forms speculative income. These operations significantly weaken the exchange rate of the national currency [8].

A specific factor of influence is the technical analysis of the foreign exchange market, which is based on predicting the exchange rate based on the quantitative analysis of available factors. Studying data on previous currency

quotes allows us to identify certain patterns of currency formation and therefore show probable changes in exchange rates in the future, both in terms of their directions, volumes, and speed. According to this concept, predicting future levels of currency quotes depends on their dynamics in the past [4].

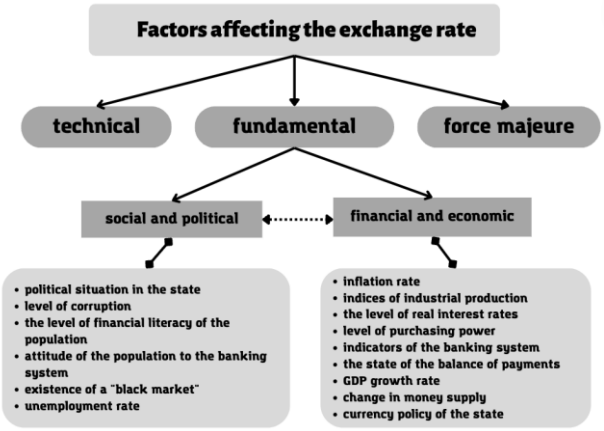


Fig. 1. Visualization of factors affecting the exchange rate

The components of the financial-economic factors influencing the exchange rate of the national currency include the main macroeconomic indicators of the country's economy, such as the inflation rate, GDP, dynamics of the money supply, various production indices, and many other indicators.

In addition to fundamental and technical factors, the impact of which can be somewhat predictable, there are also force majeure factors that can make significant adjustments to the dynamics of the exchange rate. Such force majeure factors include wars, outbreaks of epidemics, unforeseen financial and economic crises, natural disasters, technological catastrophes, and so on. All these factors have a negative impact on the stability of the national currency, as overcoming any problem requires significant resources.

To build a model for forecasting the exchange rate, the programming language Python was used. Python is a high-level object-oriented programming language used for web application development, software development, and machine learning. The Python software is presented in the form of models, which can be assembled into packages [19].

The Scikit-learn library was used for direct data analysis. It is a free machine learning software library for the Python programming language that provides functionality for creating and training various classification, regression, and clustering algorithms, and works in conjunction with NumPy. NumPy is an extension of the Python language that adds support for large multi-dimensional arrays and matrices. The Pandas software library was used for data manipulation. The Matplotlib and Seaborn libraries were used for two-dimensional or three-dimensional data visualization, providing capabilities for building graphs, scatter plots, bar and pie charts, as well as animated images.

To build a regression model for predicting the exchange rate of the Ukrainian hryvnia, data was collected on various financial and economic factors influencing the exchange rate over the period from 2010 to September 2022. Specifically, the following features were used (the encrypted name of the feature is indicated in parentheses, which is further used in the captions of the graphs and diagrams):

- Producer Price Index (PPI) – an indicator of the average level of wholesale price changes for raw materials, materials, and intermediate goods sold by national producers;
- The Inflation Index, or Consumer Price Index (CPI) – is an indicator that characterizes changes in the overall price level of goods and services that the population buys for non-production consumption. The model also uses the Consumer Price Index for the corresponding month of the previous year (inflation_p) and the Consumer Price Index for December of the previous year (inflation_12);
- Foreign Exchange Reserves (FX Reserves) – external highly liquid assets under the supervision of the state (the National Bank of Ukraine and the Government of Ukraine);
- Gross Domestic Product per capita in US dollars (gdp_pers_usd);
- Unemployment rate (unemployment) – a quantitative indicator that is determined as the ratio of the number of unemployed to the total number of the economically active working population of the country (region, social group) and is measured in percentage;
- Consolidated Balance of Payments in US dollars (BOP) – a statistical report that provides systematic information on the external economic operations of the country's residents with non-residents for a certain period;
- Real Wage Index (RSI) – an indicator that characterizes the change in the purchasing power of nominal wages;
- Net Foreign Direct Investment (FDI) balance;
- State Budget performance balance (gov_budg) or budget deficit;
- Total State Debt in US dollars (state_debt_usd);
- Balance of External Trade (int_trade) – the difference between exports and imports.

The data was collected from open sources, namely the website of the State Statistics Service of Ukraine [20] and the website of the Ministry of Finance [21]. The model also takes into account the factor of currency sales volume on the interbank currency market of Ukraine ('interbank'), data taken from the official website of the National Bank of Ukraine [22]. Since information on currency sales volumes on the interbank market is provided daily, it was summed up on a monthly basis to add to the table of data on other factors.

The official exchange rate of the Ukrainian hryvnia to the US dollar established by the National Bank of Ukraine was taken as the resulting variable. Data on the exchange rate is also provided daily, so it was aggregated monthly using a built-in function in Python. The dynamics of the Ukrainian hryvnia exchange rate for the studied period are presented in Fig. 2.

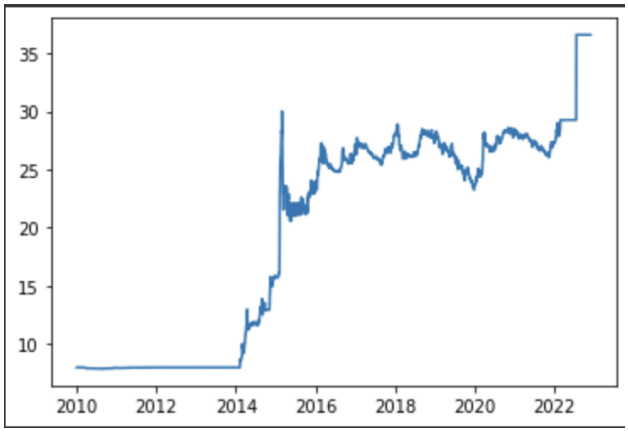


Fig.2. The official exchange rate of the Ukrainian hryvnia to the US dollar

After collecting the data and importing it into the Python programming environment, data processing was carried out. Since some data, such as GDP per capita or unemployment rate, is provided annually, while the aggregated balance of payments and the direct foreign investment balance are provided quarterly, there were many "empty" values in this data set. To solve this task, polynomial interpolation was used. The peculiarity of this type of interpolation is the construction of a polynomial P_n (x) of a degree less than or equal to n, which takes values of $f(x_i)$ at the interpolation nodes x_0, x_1, ..., x_n. The system of equations that determines the coefficients of such a polynomial has the form:

$$P_n(x_i) = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_n x_i^n = f(x_i), i = 0,1, \dots, n \quad (1)$$

The statistics for the numerical columns (count, mean, standard deviation, minimum, maximum, and quantiles) are shown in Figure 3.

	ppi	fx_reserves	gdp_pers_usd	unemployment	bop	rsi	fdi	gov_budg	state_debt_usd	int_trade	inflation_12	inflation_p	inflation	exrate	interbank
count	124.000000	124.000000	124.000000	124.000000	124.000000	124.000000	124.000000	124.000000	124.000000	124.000000	124.000000	124.000000	124.000000	124.000000	124.000000
mean	101.457443	20446.390645	3298.969816	10.205075	-213.830798	100.316129	670.183490	-36924.237903	76971.708811	-137975.599088	107.847581	110.212903	101.052419	21.934865	10186.924758
std	2.678052	6277.286724	733.444726	4.226107	1944.509634	6.808834	657.282772	73617.475860	11625.949094	85936.523909	10.522144	11.277577	1.865888	7.783870	8270.621035
min	96.200000	5625.310000	2115.000000	7.600000	-8360.000000	80.300000	-1553.000000	-418558.700000	60076.333333	-311183.000000	99.400000	100.000000	98.700000	7.990777	2858.050000
25%	100.000000	15533.125000	2641.563699	8.892778	-640.122283	97.300000	384.379348	-43683.100000	68696.843750	-182669.520492	101.300000	103.600000	100.000000	15.404735	5045.502500
50%	101.250000	20318.085000	3490.625479	9.600000	243.993651	100.400000	626.228261	-20575.550000	74815.431667	-113082.358904	104.250000	107.400000	100.800000	26.038961	6735.270000
75%	102.600000	25805.565000	3856.354795	9.970556	853.029891	103.175000	1000.420330	-2317.425000	82945.761042	-66695.943836	109.500000	109.425000	101.325000	27.181653	10860.962500
max	112.500000	31614.070000	4834.300000	35.000000	3782.000000	121.500000	2386.000000	38470.300000	123612.517241	-39475.000000	143.300000	147.100000	114.000000	36.568600	45147.060000

Fig. 3. The statistics for the numerical columns of the input data

In Figure 4, a correlation matrix of relationships between factors is presented (warmer colors indicate stronger relationships). By analyzing this matrix, we can observe the correlation between factors and the resulting feature. There is a strong negative correlation between the exchange rate and the volume of currency sales in the interbank foreign exchange market of Ukraine. It is also worth noting a fairly strong relationship between the resulting feature and the level of government debt. The exchange rate has a moderate correlation with factors such as the unemployment rate and foreign trade balance (both negative).

	ppi	fx_reserves	gdp_pers_usd	unemployment	bop	rsi	fdi	gov_budg	state_debt_usd	int_trade	inflation_12	inflation_p	inflation	exrate	interbank
ppi	1.000000	-0.036768	0.036742	-0.000524	-0.126987	-0.209618	-0.074206	0.042486	0.141598	0.127140	0.041958	0.107166	0.365768	0.189759	-0.211721
fx_reserves	-0.036768	1.000000	0.778584	0.098460	0.187987	0.046797	0.179845	-0.241109	0.506477	0.215893	-0.496570	-0.608674	-0.349499	0.022933	0.340619
gdp_pers_usd	0.036742	0.778584	1.000000	0.098914	-0.159904	-0.054161	0.105592	-0.266671	0.493679	0.512438	-0.374167	-0.462272	-0.122959	-0.255122	0.528621
unemployment	-0.000524	0.098460	0.098914	1.000000	-0.472586	-0.158106	-0.250929	-0.851443	0.663968	0.015726	0.205235	0.154151	0.158348	0.372680	-0.274529
bop	-0.126987	0.187987	-0.159904	-0.472586	1.000000	0.215657	0.146087	0.390399	-0.151850	-0.296373	-0.108229	-0.162563	-0.360512	0.161601	-0.029693
rsi	-0.209618	0.046797	-0.054161	-0.158106	0.215657	1.000000	0.095579	-0.088620	-0.091274	-0.086059	0.098253	-0.085316	-0.216852	-0.000203	0.026016
fdi	-0.074206	0.179845	0.105592	-0.250929	0.146087	0.095579	1.000000	0.217265	-0.255498	-0.031917	-0.053450	-0.021984	-0.226006	-0.280802	0.357310
gov_budg	0.042486	-0.241109	-0.266671	-0.851443	0.390399	-0.088620	0.217265	1.000000	-0.636287	-0.187391	-0.113721	0.014013	-0.044955	-0.229162	0.062365
state_debt_usd	0.141598	0.506477	0.493679	0.663968	-0.151850	-0.091274	-0.255498	-0.636287	1.000000	0.091418	-0.065932	-0.138766	0.018425	0.619703	-0.260376
int_trade	0.127140	0.215893	0.512438	0.015726	-0.296373	-0.086059	-0.031917	-0.187391	0.091418	1.000000	0.097168	0.074709	0.134878	-0.360796	0.340220
inflation_12	0.041958	-0.496570	-0.374167	0.205235	-0.108229	0.098253	-0.053450	-0.113721	-0.065932	0.097168	1.000000	0.863135	0.323193	0.137007	-0.367634
inflation_p	0.107166	-0.608674	-0.462272	0.154151	-0.162563	-0.085316	-0.021984	0.014013	-0.138766	0.074709	0.863135	1.000000	0.440099	0.193653	-0.428233
inflation	0.365768	-0.349499	-0.122959	0.158348	-0.360512	-0.216852	-0.226006	-0.044955	0.018425	0.134878	0.323193	0.440099	1.000000	0.121625	-0.266260
exrate	0.189759	0.022933	-0.255122	0.372680	0.161601	-0.000203	-0.280802	-0.229162	0.619703	-0.360796	0.137007	0.193653	0.121625	1.000000	-0.774988
interbank	-0.211721	0.340619	0.528621	-0.274529	-0.029693	0.026016	0.357310	0.062365	-0.260376	0.340220	-0.367634	-0.428233	-0.266260	-0.774988	1.000000

Fig. 4. The correlation matrix of relationships between the factors

To ensure that machine learning algorithms can work correctly with data, it is necessary to normalize them. Data normalization can be done simply by scaling them into a certain range (usually from 0 to 1), if their distribution is similar to a Gaussian distribution. In cases where the data is not normally distributed, normalization is advisable. A normal distribution of data improves the numerical stability of the model and can speed up the model training process [23].

Using the Seaborn library for data visualization, histograms were constructed for each feature to look at the data distribution. For features that did not have a bell-shaped curve distribution, normalization was performed using the Box-Cox power transformation method [24]. Examples of normalization for some of the factor variables are demonstrated in Figures 5-7.

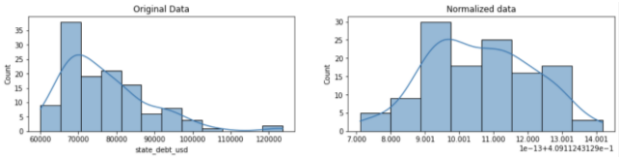


Fig. 5. Normalization of data distribution for the "state_debt_usd" indicator

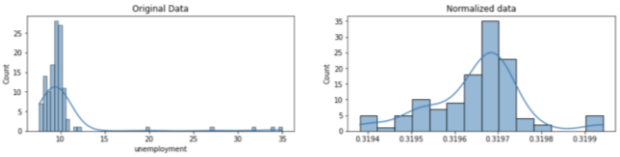


Fig. 6. Normalization of data distribution for the “unemployment” indicator

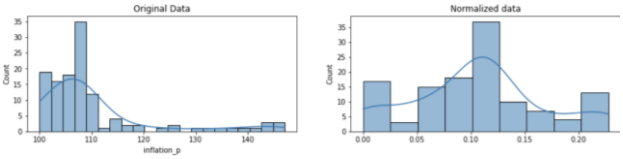


Fig. 7. Normalization of data distribution for the “inflation_p” indicator

For other features that have a similar bell-shaped distribution, standardization was performed to bring the scale of the input data into one range. This task was performed using the MinMaxScaler function (formula 2), which scaled the data into a range of 0 to 1.

$$X = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

To decode the data later, two separate scalers were used for X (predictor variables) and Y (target variable).

The next step in data processing was the detection and handling of outliers, as they can negatively affect statistical analysis and the process of training a machine learning algorithm, leading to a decrease in accuracy. An outlier is an observation in the data set that is far from the rest of the observations. This means that the outlier is significantly larger or smaller than the other values in the set. Outliers for the given features can be seen in the histograms of the data distribution. Outliers can also be conveniently identified using box plots.

Figure 8 shows box plots for four indicators, with points that are outliers and do not fall within the range of other observations, meaning they are not close to the quartiles. To remove these outliers, a function was written that sets them equal to a certain quartile, which is manually selected.

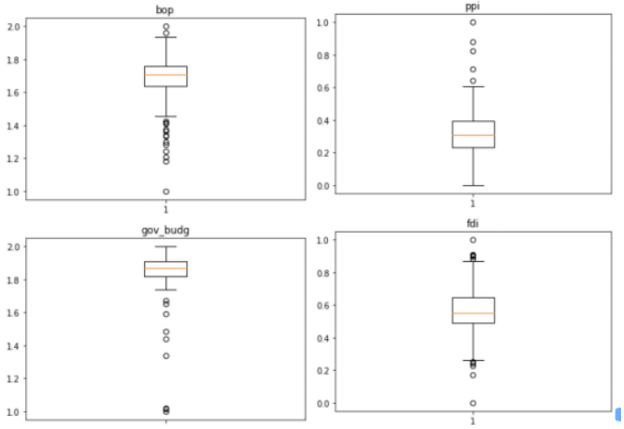


Fig. 8. Visualization of the distribution of variable values using box plots for four factors

Figure 9 shows box plots of the same factor variables, but after outliers have been corrected. As can be seen, there is no longer such a strong data spread.

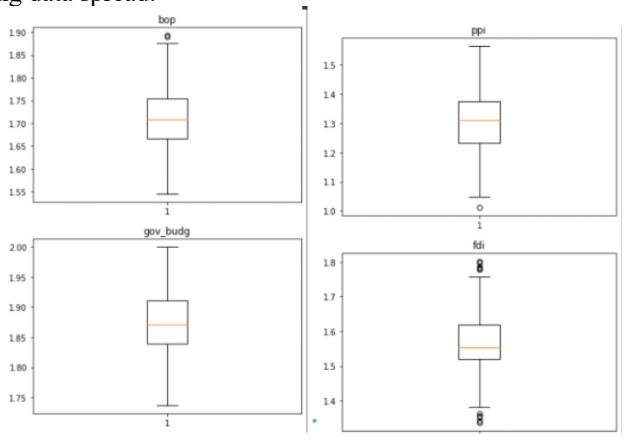


Fig. 9. Visualization of the distribution of variable values using box plots after outlier correction

The model was built using the Extreme Gradient Boosting (XGBoost) method, which is a scalable machine learning library with distributed decision trees and gradient boosting. It is a leading machine learning library for regression, classification, and ranking tasks. Decision trees create a model that predicts a label by evaluating decision tree questions about the if-then-else function, true/false, and evaluating the minimum number of questions needed to estimate the probability of making the correct decision. Gradient Boosted Decision Trees (GBDT) is a decision tree ensemble learning algorithm similar to random forests for classification and regression. Ensemble learning algorithms combine multiple machine learning algorithms to obtain a better model. Gradient boosting is an extension of boosting, where the process of additive generation of weak models is formalized as a gradient descent algorithm over the target function. Gradient boosting sets target outcomes for the next model to minimize errors. The target outcomes for each case are based on the error gradient (hence the name gradient boosting) concerning the prediction. The final prediction is the weighted sum of all tree predictions.

XGBoost is a scalable and high-precision implementation of gradient boosting that extends the boundaries of computational power for enhanced tree-like algorithms, mainly designed to improve the productivity of machine learning models and computation speed. With XGBoost, trees are built in parallel, adhering to a level-wise strategy, scanning gradient values, and using these partial sums to evaluate the splitting quality at each possible split in the training set.

Modeling with XGBoost begins with model training, which was conducted based on 86% of the input data. Mean absolute error and mean squared error functions were used to evaluate the model's adequacy.

The mean_absolute_error function calculates the average absolute error, a risk metric that corresponds to the expected value of the absolute error or loss norm.

$$MAE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i|, \quad (3)$$

where $\hat{y}_i$ – is the predicting value of i-sample and y_i – is the corresponding truth value.

The mean_squared_error function calculates the mean squared error, a risk metric corresponding to the expected value of the squared error or loss:

$$MSE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2, \quad (4)$$

where $\hat{y}_i$ – is the predicting value of i-sample and y_i – is the corresponding truth value.

As a result of constructing a gradient boosting model, the following results were obtained: the MAE and MSE indicators are 0.0560 and 0.0100, respectively. The prediction errors turned out to be quite low, which means that the model adequately predicts the exchange rate of the Ukrainian hryvnia given the specified factors.

The next stage of our research was to determine the importance of the selected factors on the exchange rate of the Ukrainian hryvnia. Figure 10 shows the factor importance indices for the model. As can be seen, the most important factor for the exchange rate of the Ukrainian hryvnia, based on this model, is the level of inflation, specifically the consumer price index for the corresponding month of the previous year (the importance coefficient is close to 0.5). Inflationary processes in the country lead to a decrease in purchasing power and a tendency for the national currency to fall against currencies in countries where inflation is lower. The factor of GDP per capita also has a significant impact: the higher the GDP growth rate, the higher the demand for the national currency and therefore a higher exchange rate. The next factors, whose importance coefficients exceed 0.1, are the unemployment rate and the country's government debt; an increase in these indicators has a negative impact on the national economy and, accordingly, on the exchange rate of the national currency. The volume of gold and foreign exchange reserves also plays a role in determining the exchange rate, if a country does not have sufficient resources to support the exchange rate, it becomes more vulnerable to speculative attacks. Additionally, the factor of international trade to some extent determines the demand for the national currency. The influence of other factors included in this model is somewhat less significant.

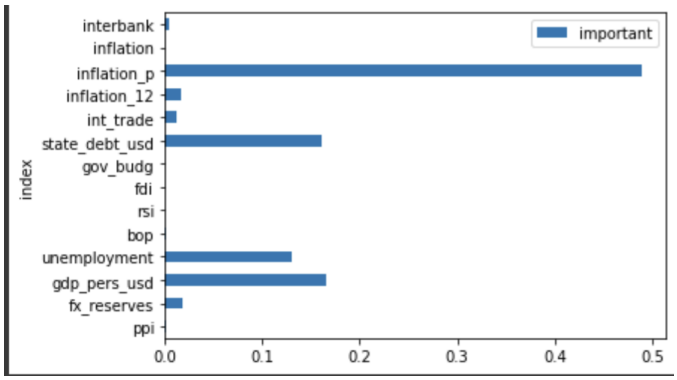


Fig.10. Coefficients of the importance of factors for the model

The final stage of our calculations is to make a forecast for the exchange rate of the Ukrainian hryvnia. Figure 11 shows retrospective forecast values of the exchange rate of the Ukrainian hryvnia for comparison with the real values. The real exchange rate is represented by the blue color on the diagram, while the forecast is represented by the red color.

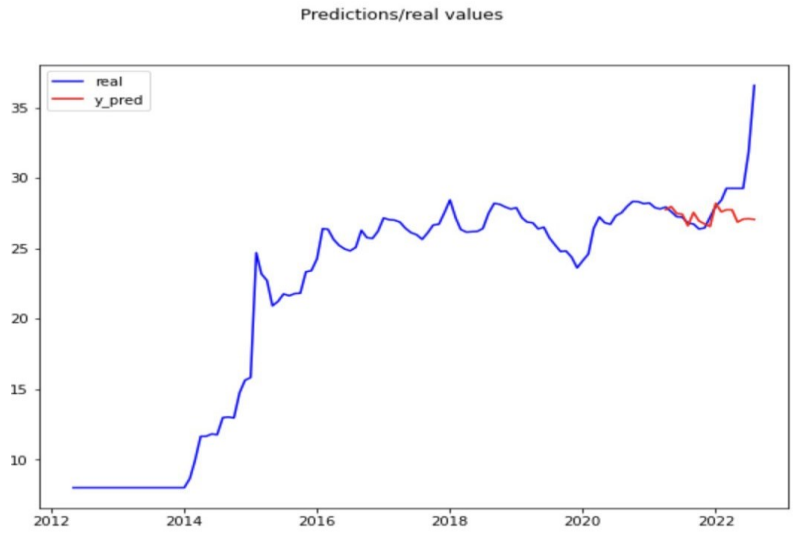


Fig. 11. Chart of the Ukrainian hryvnia exchange rate and retrospective chart of the forecast

Analyzing the chart, we can conclude that in 2021 the retrospective forecast corresponded to the main trends and moved in parallel directions with the real exchange rate, indicating the adequacy of the model. However, in 2022, force majeure factors came to the fore, which cannot be predicted, and even if possible, the impact on the economy of the country and the exchange rate, including the economic-mathematical models used, cannot be estimated. As of the beginning of 2022, the financial and economic indicators of Ukraine did not show negative trends that could have caused an increase in the exchange rate of the Ukrainian hryvnia against the US dollar, but the Russian invasion and the beginning of full-scale war caused a colossal destructive impact on the economy of Ukraine and the national currency exchange rate in particular. Since the model is trained on historical data that did not contain such unprecedented events, or so-called "black swans," the forecast data obtained from our calculations do not correspond to the actual data. At this stage of computer modeling development, it is still difficult to predict such incidents, let alone the impact of such extraordinary events on the economy of the country and the exchange rate of the national currency in particular.

Conclusions

As studies have shown, it is possible to analyze, model, and forecast the exchange rate of a national currency in a country without force majeure circumstances using machine learning methods based on a pre-built multifactor regression dependence of this rate on macroeconomic factors. The results of the retrospective forecast of the Ukrainian hryvnia exchange rate confirmed the high accuracy and effectiveness of the proposed method of forecasting this rate. In 2021, the retrospective forecast corresponded to the main trends and moved in parallel directions with the real rate, indicating the adequacy of the developed model. Under stable political conditions and projected socio-economic development, this model is likely to predict certain fluctuations in the Ukrainian hryvnia exchange rate. The scientific results obtained in the work regarding the proposed approaches to forecasting the national currency exchange rate using machine learning methods should be used in practice by all interested parties, including financial institutions of Ukraine, to achieve stability of the national currency, which in turn will affect the development of the national economy as a whole and the welfare of the population of the country. In the process of further research, the proposed approach to forecasting the country's exchange rate can be improved by adding additional factors that affect this rate to the considered model.

References

1. Bakumenko T.V. Exchange rate and fundamental factors of its formation. *Problems and prospects of development of the banking system of Ukraine: coll. of science works*. Sumy: UAB NBU, 2004. Vol. 9. P. 344-352. URL: https://essuir.sumdu.edu.ua/bitstream-download/123456789/54716/5/Bakumenko_Valiutnyi_kurs.pdf (Accessed on: 20.12.2022).
2. Malashchuk D.V. Analysis of currency exchange rate formation factors. *Foreign trade: economics, finance, law*. 2012. No. 6. P. 83-86. URL: http://nbuv.gov.ua/UJRN/uazt_2012_6_16 (Accessed on: 12.12.2022).
3. Chirka D. M. Exchange rate and its influence on the activity of business entities. *Visnyk ZSTU. Economic sciences*. 2010. No. 3 (53). P. 201-203.
4. Dzyublyuk O.V. Banking operations: Textbook. Ternopil: «Economic Thought» Publishing House of TNEU, 2009. 696 p.
5. Kozlovskyi S.V., Kozlovskyi V.O. Macroeconomic modeling and forecasting of the exchange rate in Ukraine: Monograph. Vinnytsia: «Vega Book» JSC Vinnytsia Regional Printing House. 2005. 240 p.
6. Marchenko V.M. Factors of exchange rate changes in Ukraine. *Contemporary problems of economy and entrepreneurship*. Issue 19. Kyiv: Publishing house of National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute». 2017. P 59-66.
7. Savchenko T.G., Yepifanova M.A. Evaluation of the effectiveness of instruments of currency regulation in Ukraine. Actual economic problems. 2011. No. 2. P. 161-170.
8. Korneeva Yu. V. The choice of the currency regime and the effectiveness of monetary policy in the transition economies of the CEE countries. *Economic space*. 2010. No. 36. P. 13-24.
9. Drebot N.P. Factors influencing the exchange rate of the national currency in Ukraine. Scientific Bulletin of National Technical University of Ukraine. 2016. Issue 26.2. P. 190-196.
10. Krukovets D. Data Science Opportunities at Central Banks: Overview. Visnyk of the National Bank of Ukraine, 2020. No. 249. P. 14-26. https://doi.org/10.26531/vnbu2020.249.02.
11. Rizk A., Elragal A. Data science: developing theoretical contributions in information systems via text analytics. Journal of Big Data. No.7 (2020). https://doi.org/10.1186/s40537-019-0280-6.
12. Duggal N. The best introduction to Data Science. URL: https://www.simplilearn.com/tutorials/data-science-tutorial/introduction-to-data-science (Accessed on: 24.11.2022).
13. Hui Lin, Ming Li. Introduction to Data Science. URL: https://scientistcafe.com/ids (Accessed on: 03.12.2022).
14. Ensembles of machine learning models. URL: https://evergreens.com.ua/ua/articles/ensembles.html (Accessed on: 08.11.2022).
15. Learn Types of Machine Learning Algorithms with Ultimate Use Cases. URL: https://data-flair.training/blogs/types-of-machine-learning-algorithms (Accessed on: 11.01.2023).
16. Using XGBoost in Python Tutorial. URL: https://www.datacamp.com/tutorial/xgboost-in-python (Accessed on: 15.02.2023).
17. How to develop your first XGBoost Model in Python. URL: https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn (Accessed on: 04.01.2023).
18. Time-Series Analysis guide. URL: https://www.kaggle.com/code/andreshg/timeseries-analysis-a-complete-guide (Accessed on: 29.11.2022).
19. What is Python programming language. URL: https://freehost.com.ua/ukr/faq/wiki/chto-takoe-jazik-programmirovanija-python (Accessed on: 17.12.2022).
20. Website of the State Statistics Service of Ukraine. URL: https://www.ukrstat.gov.ua (Accessed on: 14.11.2022).
21. Website of the Ministry of Finance. URL: https://minfin.com.ua (Accessed on: 19.11.2022).
22. Official website of the National Bank of Ukraine. URL: https://bank.gov.ua (Accessed on: 22.11.2022).
23. Understand Data Normalization in Machine Learning. URL: https://towardsdatascience.com/understand-data-normalization-in-machine-learning-8ff3062101f0 (Accessed on: 12.12.2022).

24. Box Cox Transformation: Definition, Examples. URL: https://www.statisticshowto.com/probability-and-statistics/normal-distributions/box-cox-transformation (Accessed on: 23.12.2022).

Vasyl Pryimak **Василь Приймак**	Doctor of Economic Sciences, Professor, Head of Management Information Systems department, Ivan Franko National University of Lviv, Ukraine, e-mail: vasyl.pryymak@lnu.edu.ua https://orcid.org/0000-0003-0244-8661	доктор економічних наук, професор, завідувач кафедри інформаційних систем у менеджменті, Львівський національний університет імені Івана Франка, Львів, Україна
Bohdan Bartkiv **Богдан Бартків**	MSc student of Management Information Systems department, Ivan Franko National University of Lviv, Ukraine, e-mail: bohdan.bartkiv@lnu.edu.ua https://orcid.org/0009-0001-2061-6676	магістрант кафедри інформаційних систем у менеджменті, Львівський національний університет імені Івана Франка, Львів, Україна
Olga Holubnyk **Ольга Голубник**	PhD, Associate Professor of Management Information Systems department, Ivan Franko National University of Lviv, Ukraine, e-mail: olga.holubnyk@lnu.edu.ua https://orcid.org/0000-0003-1211-4614	кандидат економічних наук, доцент кафедри інформаційних систем у менеджменті, Львівський національний університет імені Івана Франка, Львів, Україна

UDC 004.9
https://doi.org/10.31891/csit-2023-1-11

Khrystyna ZUB, Pavlo ZHEZHNYCH
Lviv Polytechnic National University

MACHINE LEARNING BOOSTING METHODS FOR PREDICTION A HIGHER EDUCATION INSTITUTIONS ENTRANT'S ADMISSIONS IN UKRAINE

There is a constant and growing need for higher education institutions (HEI) to provide proper and high-quality support for the admissions campaign through information systems and technologies. Labor market trends, unreliability and low-quality sources, and a large volume of admission rules can complicate the admission process for an applicant. As a result, there is a risk that the applicant will not be able to make the right choice and quality assessment of the chances of admission. So, this paper considers increasing the entrant's chances of making an effective decision at the stage of education program selection through the implementation of an information system. The efficiency of such systems is largely based on the accuracy of its intelligent components. This article investigates the effectiveness of machine learning (ML) boosting methods to solve the problem of admission prediction through binary classification tasks. We evaluate the accuracy of such ML methods as Gradient Boosting, Adaptive Boosting (AdaBoost), and eXtreme Gradient Boosting (XGBoost). For a more detailed assessment of the studied methods, a comparison with Support Vector Machine (SVM) and Logistic regression is also presented. The simulation was performed using «Orange» software. The work of the studied methods was simulated based on a sample of archival data comprising 9,657 records of full-time entrants of two faculties of Lviv Polytechnic National University. The sample was randomly divided into training and test sets in a ratio of 80% to 20%. To ensure the reliability of the obtained result, the work of each of the studied methods was subjected to 10-fold cross-validation. Classification accuracy (AUC), Precision, Recall and F1 score performace indicators was used to analyze the results. It has been experimentally established that the highest accuracy is achieved when using XGBoost. The obtained results shows high accurate. This makes it possible to use the researched methods in the subsequent stages of building an information system to support the decision-making of applicants.

Keywords: admission, entrant, higher education institution (HEI), prediction, machine learning, boosting, information system.

Христина ЗУБ, Павло ЖЕЖНИЧ
Національний університет «Львівська політехніка»

БУСТИНГОВІ МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ УСПІШНОСТІ ВСТУПУ АБІТУРІЄНТІВ ЗВО УКРАЇНИ

Існує постійна та зростаюча потреба закладів вищої освіти (ЗВО) у забезпеченні належного та якісного супроводу вступної кампанії за допомогою інформаційних систем та технологій. Тенденції на ринку праці, ненадійність і неякісність джерел), велика кількість правил прийому можуть ускладнити процес вступу абітурієнта. Як наслідок, є ризик того, що абітурієнт не зможе зробити правильний вибір та якісно оцінити шанси на вступ. Тож, у даній роботі розглядається завдання підвищення шансів абітурієнта прийняти ефективне рішення на етапі вибору освітньої програми. Ефективність таких систем значною мірою базується на точності їх інтелектуальних компонентів. У цій статті досліджується ефективність бустингових методів машинного навчання для вирішення проблеми прогнозування вступу за допомогою завдань бінарної класифікації. Ми оцінюємо такі точність роботи таких методів машинного навчання, як Gradient Boosting, Adaptive Boosting (AdaBoost) і eXtreme Gradient Boosting (XGBoost). Для більш детальної оцінки досліджуваних методів також представлено порівняння з методом опорних векторнів і логістичною регресією. Моделювання проводилось за допомогою програмного забезпечення «Orange». Роботу досліджуваних методиів було змодельовано на основі вибірки архівних даних, яка склала 9657 записів даних абітурієнтів денної форми навчання двох навчально-наукових інститутів Національного університету «Львівська політехніка». Вибірку випадковим чином було розподілено на навчальну та тестову вибірки у співвідношенні 80% до 20%. Для забезпечення достовірності отриманого результату роботу кожного з досліджуваних методів піддавали 10-кратній крос-валідації. Для аналізу результатів використано такі показники точності як Classification accuracy (AUC), Precision, Recall, F1 score. Експериментально встановлено, що найвища точність досягається при використанні XGBoost. Отримані результати досить точні. Це дає можливість використовувати досліджувані методи на наступних етапах побудови інформаційної системи підтримки прийняття рішень абітурієнтами.

Ключові слова: вступ, абітурієнт, заклад вищої освіти (ЗВО), прогнозування, машинне навчання, бустинг, інформаційна система.

Introduction

The rapid development of information technologies, as well as means of artificial intelligence, contributes to their wide application, in particular in the field of education. In recent years ML has found larger and broader applications in HEI and is showing an increasing trend in scientific research that considers the increasing effectiveness of entrants' admission.

As the admission results directly affect young workers' professional trajectories, it is important to provide appropriate support for entrants at this stage. The world markets are developing rapidly and continuously looking for the best knowledge and experience among people. HEI rating, a wide range of educational programs, a constantly changing labor market, admission rules, incompetent recommendations, or career guidance activities could make this decision complicated. As a result, they could make unwise choices. To increase the entrant's chances of making an effective decision, this research aims to investigate the possibility of using ML techniques to predict their chances of admission.

HEI admission process could differ in all countries and requires investigation in each individual case. This research considers the Ukrainian HEI entrant's decision. Today, Ukrainian HEI provides the target audience with up-to-date and reliable information on official websites in a large volume and an accessible form. There are also open online services that provide additional information about rating lists and the education programs of admission. But still, there is a need to provide any free and open web resource that could support entrant's admission decisions.

According to the admission rules, everyone who intends to admit the HEI must register and pass an external independent assessment test aimed at determining the level of educational achievements of graduates of secondary educational institutions upon their admission to HEI in Ukraine. It is worth noting that the institution sets a minimum passing score for specific education programs. There is also a limit on the number of submitted applications. In every application submitted for the budget form of study, it is necessary to set its priority - from one to five, where one is the highest priority and five is the lowest. After submitting the first application, it will not be possible to change the priority. In accordance with the recommendations of the HEI, which has to be conducted based on the competition, rating lists are published, after which the entrant makes the final choice.

Such a complex admission system is aimed at raising the level of education of the population of Ukraine and ensuring the implementation of the constitutional rights of citizens to equal access to higher education, monitoring compliance with education standards, analyzing the state of the education system, and forecasting its development. However, at the same time, it is difficult for school graduates. There is no clear predictive vision of the most appropriate specialization for the student.

The wrong decision causes the entrant not to get state financing, enter to undesirable program, and could become a not success d student. Given the difference in the admission procedure of foreign higher education institutions, Ukraine cannot adopt the experience of supporting entrants. However, taking into account modern trends and existing solutions, in this study we will analyze the relevance of using ML methods on the example of an admission campaign of Lviv Polytechnic National University admission data.

The aim of the work is to apply and experimentally evaluate the accuracy of ML algorithms in solving the task of predicting the chances of admission to the HEI. The performance indicators will be consider as an indicator of model effectiveness. The most effective model could be used as a base intellectual component for an information system aimed to support Ukrainian entrants.

Related works

A huge amount of students and entrants data is available in many HEI, which may be utilized to make future decisions and improve future outcomes. As a result, there is a growing number of scientific research using ML methods in the education area tasks [1]. Fig. 1 shows the growing number of researches in the Scopus database related to the use of ML in HEI. This section summarizes the main directions and obtained results of existing latest studies aimed to predict HEI admission and describes some of them.

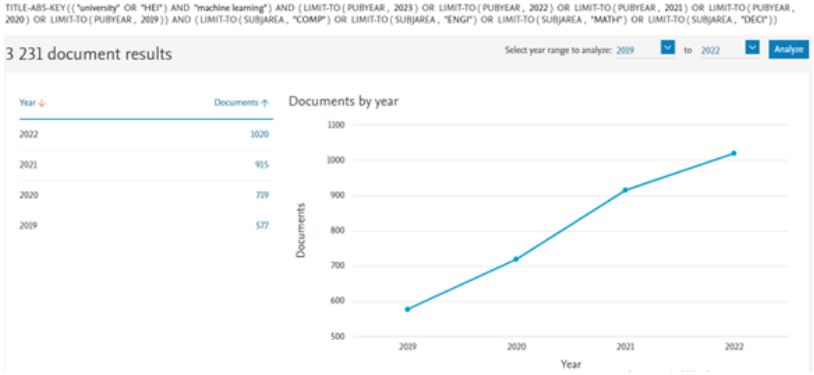


Fig.1 Scopus search analysis dashboard

A common practice is to compare ML methods for solving classification problems. Applying such methods as Logistic Regression, KNN Classification, SVM, Naive Bayes Classification, Decision Tree Classification, and Random Forest authors aimed to predict the admission outcome of candidates with a set of known parameters. Comparing the performance metrics of these methods allowed to highlight the most effective solution for each data set [2-6].

As one of effective techniques authors used stacking to predict admission to a bachelor's program. It performs better while compared to other regression algorithms such us, Linear Regression, SVM, Decision Tree, Random Forest, and Stacking Regressor. To analyze all the models they presented evaluation metrics such as R2 Score, Mean Absolute Error, Mean Square Error, and Root Mean Square Error for each case [7].

In other paper, authors use ML methods to resolve classification task to identify the possibility of enrollment for a pool of applicants. There are two approaches presented in the study. The first one, based on SVM and LR models, uses a given set of features and defines the total number of enrollments. The second approach directly implements a semi-supervised probability model and a time series model and determines the number of applicants without

identifying them individually. The results demonstrate that the presented models can predict enrollment with reliable accuracy using only a small set of features related to student and college features [8].

With the aim to rid of the accuracy limitation produced by existing web applications or consultancy services, this study uses a deep neural network (DNN) to predict the chance of getting admitted to a university according to the student's portfolio. The DNN model outperformed the results in comparing with existing methods in terms of different performance metrics in each benchmark [9].

Also, in one of the recent researches, the authors emphasize that there is a high level of unemployment and the absence of enough internship or full-time job opportunities for college graduates caused of the coronavirus. At the same time, the number of post-graduate applicants is growing. This makes the need to help applicants to understand their overall admission chances. They use the feasible ordinary least squared multiple linear regression models to analyze and predict the post-graduate school admission chance [10].

Current studies present experiment result based on datasets of various HEI. The problem-solving of entrants' admission chances evaluation is critical in each country. There are studies aimed to support Bangladeshi students intended to pursue higher studies abroad after completing their undergraduate degrees [11], to make recommendations for Indian students who apply for the admissions of overseas universities to study abroad [12], for foreign students came to USA [13].

They also analyze a student's academic achievements historical data to predict graduate program admission [14, 15, 16, 17.] or first-year admission [18, 19].

So today, various types of research consider the effectiveness of the application of ML methods in HEI admission procedures, particularly an admission prediction task The analyzed studies confirm the practicality and value of solving the problem of predicting admission. All the studies summarize that the presence of a decision support system will positively affect both the life trajectory of the future student and the activities of the HEI. They show high accuracy in terms of every settled task and used dataset. This set differs in individual cases according to the task at hand. But the existing studies concerned different HEI, education programs, and even countries, different levels - the first year, re-enrollment, master's, or Ph.D. It is obvious that there is a need to collect and use specific data in every separate case.

Since the research concerns the entrance of other countries and different admission rules, the task of researching the application of ML models for predicting Ukrainian applicants' admission remains open.

Dataset description

The selection of data was limited by the following characteristics: entry year - 2021, qualification level - bachelor's (entry to the first year), form of study - full-time, faculties - humanities and social sciences, computer sciences and technologies. In connection with the beginning of the Russian-Ukrainian war, the rules of reception have partially changed. As a return to typical procedures is envisaged, based on the external examination, the analysis was carried out on the basis of data from the 2021 entry campaign. Data source: vstup2021.lpnu.ua [20]. Total row number is 2057 for the first case and 5600 for the second. Data preprocessing is described below.

Exploratory data analysis shows some outliers. There are many different methods to deal with outliers - leave as is, drop them with dropna, fill missing values with test statistics like mean. In our case, there are few outliers. Therefore, the best option will be removing the rows that contain outliers - applicants who entered with low scores based on the privileges granted to them due to special cases of the admission rules. In order to removing redundancy we deleted row id because it does not correlate to the dependent variable; hence, it was removed from the dataset.

Instead of introducing separate properties that correspond to the results of the external examination for each subject, we used the Competitive score - this is a comprehensive assessment of the entrant's achievements, which includes the results of entrance tests and other indicators calculated in accordance with the rules of admission to HEI. In addition to the results of external examinations, the formula also provides for integral weighting factors determined for each specialty and additional points for successfully completing the preparatory courses of a HEI in the year of admission. In addition, corrective coefficients are also taken into account - rural (can be calculated for entrants registered in the village and who graduated from a rural school in the year of admission), regional (for those entering regional universities) and branch (for specialties that receive special support). These data are not listed separately in the data source, but they are taken into account in the competitive score.

Proposed technique

To obtain more accurate results when applying ML methods, the algorithm of their ensembles is used, which consists of the simultaneous application of several basic algorithms. When using them, the algorithms learn simultaneously and can correct each other's errors. The ensemble itself is a supervised learning algorithm because it can be trained and then used to make predictions, so the trained ensemble represents a single hypothesis. This hypothesis, however, does not necessarily lie in the hypothesis space of the models from which it is constructed. Thus, ensembles can have more flexibility in the functions they represent.

There are several approaches to building ensemble algorithms, including stacking, boosting, and bagging. Boosting consists of the gradual application of each model while each subsequent one corrects the errors of the

previous one. During bagging, the basic algorithms are trained in parallel, and the final results are aggregated. In stacking, several different algorithms are trained, and at their outputs, the metamodel produces the final result.

There are a number of key advantages that the boosting method provides when used for classification or regression problems, such as ease of implementation, bias reduction, and computational efficiency. So, in this work, we investigated the accuracy of solving the binary classification task using a number of existing methods of ML, in particular: AdaBoost, XGBoosting, and Gradient Boosting. We aim to evaluate the accuracy and efficiency of boosting methods and the possibility of their future implementation. For better evaluation, we compare its result with other different ML methods, such as SVM and Logistic Regression, for two separate datasets.

AdaBoost was the first really successful boosting algorithm developed for the purpose of binary classification. AdaBoost is a very popular boosting technique that combines multiple "weak classifiers" into a single "strong classifier". Building on the work of Leo Breiman, Jerome H. Friedman developed gradient boosting, which works by sequentially adding predictors to an ensemble, with each one correcting for the errors of its predecessor. However, instead of changing the weights of data points like AdaBoost, the gradient boosting trains on the residual errors of the previous predictor. The name, gradient boosting, is used since it combines the gradient descent algorithm and boosting method.

XGboost is a decision-tree-based ensemble ML algorithm that uses a gradient boosting framework. It is one of the gradient boosting implementations that is acknowledged as one of the best-performing algorithms used for supervised learning. Its main advantage is high execution speed out of core computation. XGBoost algorithm was developed as a research project at the University of Washington. Since its introduction, this algorithm differentiates itself in a wide range of applications - can be used to solve regression, classification, ranking, and user-defined prediction problems. XGboost preferred by data scientists because its high execution speed out of core computation. More detailed description can be found in [21].

The simulations were performed using "Orange" software [22], an open-source online data visualization, ML, and data analysis tool. It is equipped with a visual programming interface for fast qualitative analysis and interactive visualization of data. The block diagram of this process is presented in Fig. 2.

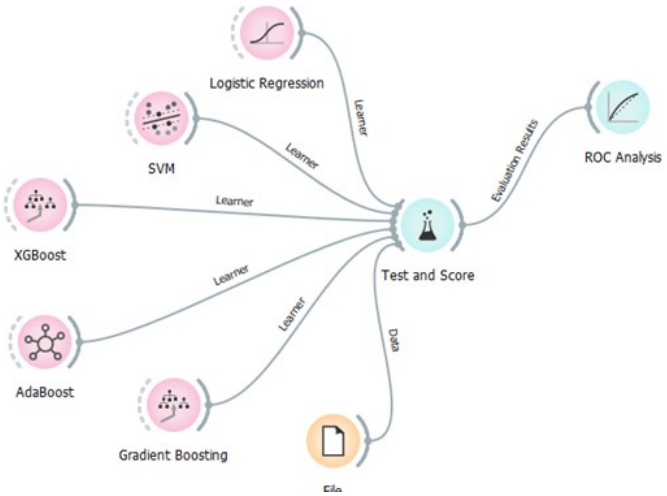


Fig. 2 Research simulation in the "Orange" software

Due to the sufficient size of the data sample, it was divided into training and test samples in the ratio of 80% to 20%. Classification accuracy was assessed using 10-fold cross-validation. The essence of such a check is to compare the results of the classification accuracy of the test and training sets. It is considered that the studied methods has passed the test, provided that the classification of the test set gives approximately the same results in terms of accuracy as the classification of the training set.

To evaluate the performance of the ML method, we use following performance indicators:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \quad (1)$$

$$Precision = \frac{TP}{TP + FP}, \quad (2)$$

$$Recall = \frac{TP}{TP + FN}, \quad (3)$$

$$F1\_score = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)}, \quad (4)$$

where: TP are true positive observations, TN are true negatives observations, FP are false positive observations, FN are false negatives observations.

Result and discussion

This section presents the results of work and a comparison of such ML methods as Gradient Boosting, AdaBoost, and XGBoost. Table 1 and Table 2 describe the results obtained for each data set according to the accuracy evaluation measures described in the previous section.

Table 1

Results of boosting ML methods, dataset 1

Model	AUC	F1	Precision	Recall
XGBoost	0.890	0.919	0.919	0.924
Gradient Boosting	0.888	0.915	0.917	0.921
AdaBoost	0.804	0.890	0.889	0.891

Table 2

Results of boosting ML methods, dataset 2

Model	AUC	F1	Precision	Recall
XGBoost	0.972	0.962	0.962	0.963
Gradient Boosting	0.967	0.962	0.961	0.962
AdaBoost	0.891	0.930	0.930	0.931

Taking into account the fact that the competitive scores of entrants of humanities majors can be higher on average than those of mathematics and natural sciences, therefore, in this study, two separate cases of different faculties are cursed and evaluated. This will allow for a more detailed evaluation of the effectiveness of the selected methods and to avoid discrimination of technical specialties of higher education institutions.

The findings of this study clearly show that the XGboost method gives the most accurate results. The obtained numerical values of the metrics were also confirmed by visual analysis. An error curve, the so-called ROC curve, was used to visualize the research results (Fig. 3, Fig. 4). This is one of the most commonly used methods of demonstrating binary classification results. The ROC curve shows the dependence of the number of true positive values on the number of false positive values for each individual class. Accordingly, the studied classifier, whose ROC curve is located above and to the left of the graph, demonstrates greater accuracy.

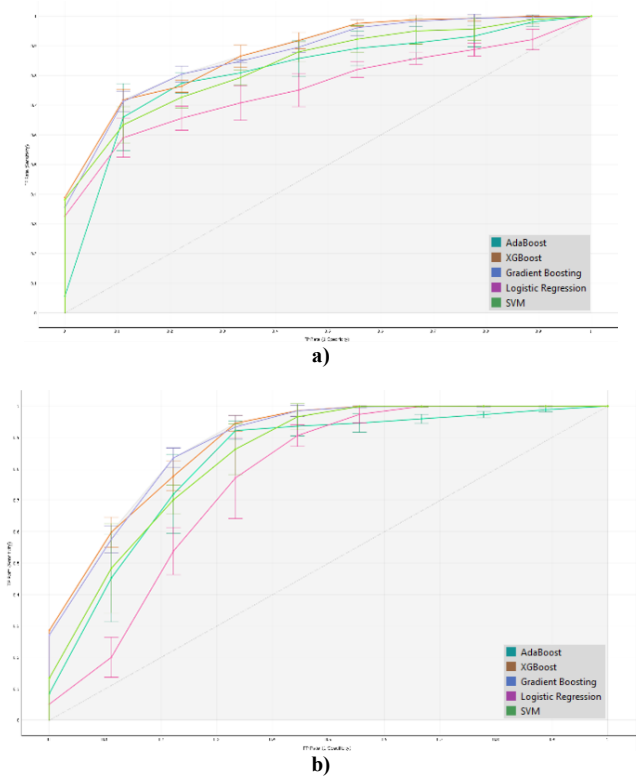
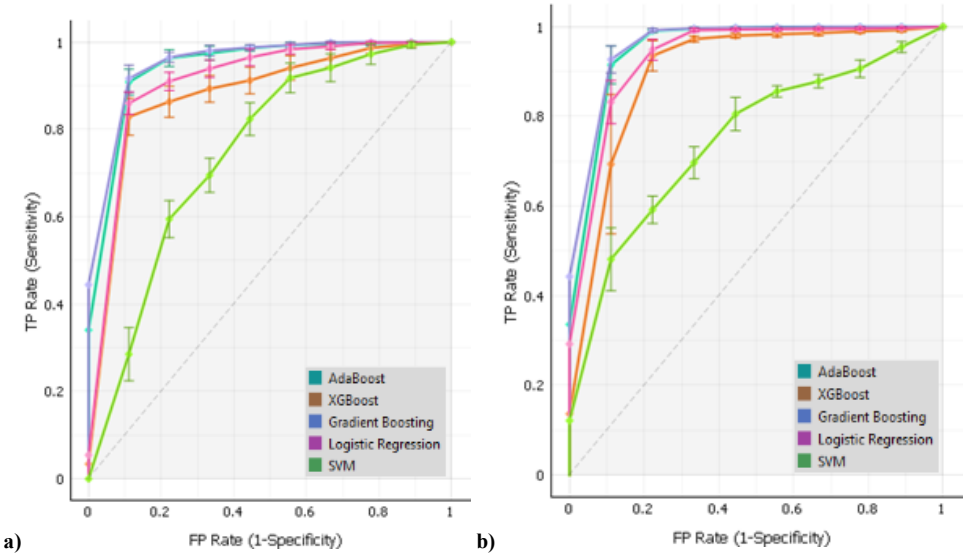


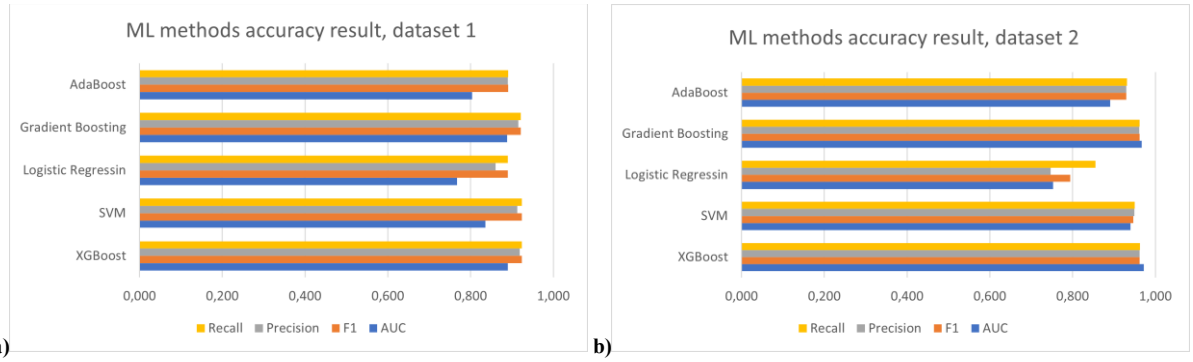
Fig. 3 ROC-curves for the boosting methods, dataset 1: a) class 1 (admitted); b) class 2 (not admitted)



a) b)

Fig. 4 ROC-curves for the boosting method, dataset 2: a) class 1 (admitted); b) class 2 (not admitted)

For a better assessment of the researched methods, an analysis of the work of such methods as SVM and Logistic Regression was also carried out. The results are presented in Figure 5.



a) b)

Fig.5 Comparison of the accuracy of the boosting methods with classical ML methods: a) dataset 1, b) dataset 2

As can be seen from both graphs of Figs. 3 and Fig. 4, ROC-curves several algorithms almost overlap. This indicates that they are approximately equally effective. This is confirmed by the results presented in Table 1, and Table 2. However, XGBoost shows slightly better results. This is also confirmed by numerical estimates of comparison with other methods presented at Fig. 5. This provides the possibility of applying this method when building a real system for predicting the success of the entrant's entry to HEI.

Conclusion

ML algorithms are widely used in many fields of scientific and practical activity, including education. Ensemble learning has been commonly used in machine learning on various classification and regression tasks to improve performance by grouping individual algorithms. Boosting is one of the most popular ensemble learning techniques, where a set of so-called weak learners, i.e., models whose performance is slightly better than random guessing, is built.

The current development of decision-making support systems for applicants requires the search and application of the most effective and accurate methods of predicting admission success. The need for the entrants to decide on the choice of university and education program for admission arises every year during the admissions campaign. So, supporting applicants remains relevant for all educational institutions. From the research, we were able to verify the effectiveness of the application of ML methods and techniques to solve such a task.

This paper examines the task of supporting applicants' decision-making to HEI choosing an education program for admission using ML boosting methods. It has been experimentally established that the highest accuracy is achieved using the XGBoost method. The obtained results make it possible to consider the boosting method as the basic algorithm of the recommender system, i.e., the components and decision support of applicants to the University of Ukraine. Although the XGBoost method showed the highest accuracy, further research will be conducted in the more detailed investigation of boosting methods and a more full dataset.

References

1. Scopus - Elsevier's abstract and citation database. URL: https://www.scopus.com (revised on 10.01.2023).
2. Iman A., Tian X. A Comparison of Classification Models in Predicting Graduate Admission Decision. *Journal of Higher Education Theory and Practice*. 2021. Vol. 21, No. 7. P. 219-230.
3. Jeganathan S., Parthasarathy S., Lakshminarayanan A. R., Ashok Kumar P. M. and Khan M. K. A. Predicting the Post Graduate Admissions using Classification Techniques. *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, Pune, India, 2021. P. 346-350.
4. Kalakova A., Amanbek Y., Kairgeldin R. and Kalakova G. Classification and prediction of Student's Enrollment to Kazakhstanis Universities Using Characteristics of Applicant and Testing Results. *IEEE International Conference on Smart Information Systems and Technologies (SIST)*. Nur-Sultan, Kazakhstan, 2021. P. 1-7.
5. Kruthika CS., Apeksha B., Chinmaya GR., Madhumathi JB., Veena MR. University Admission Prediction using Machine Learning. *Global Journal of Research and Review*. 2021. Vol. 8, No. 7.
6. Zub K., Zhezhnych P. Performance evaluation of ML-based classifiers for HEI Graduate Entrants. *International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2021),* Kharkiv, Ukraine, September 20-21, 2021. P. 92-97.
7. Ace C. Lagman, Lourwel P. Alfonso, Marie Luvett I. Goh, Jay-ar P. Lalata, Juan Paulo H. Magcuyao, and Heintjie N. Vicente. Classification Algorithm Accuracy Improvement for Student Graduation Prediction Using Ensemble Model. *International Journal of Information and Education Technology*. 2020. Vol. 10, No. 10. P/ 723-727.
8. Slim A., Hush D., Ojah T., Babbitt T. Predicting Student Enrollment Based on Student and College Characteristics. *Proceedings of the 11th International Conference on Educational Data Mining,* Raleigh, NC, Jul 16-20, 2018. P. 383-389.
9. Omaer Faruq Goni M., Matin A., Hasan T., Abu Ismail Siddique M., Jyoti O. and Sifnatul Hasnain F. M. Graduate Admission Chance Prediction Using Deep Neural Network. *IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering* (WIECON-ECE), Bhubaneswar, India, 2020. P. 259-262.
10. Bo Liu, Research on the Admission of Graduate Students Based on Multiple Regression Model. *Proceedings of the 2022 International Conference on mathematical statistics and economic analysis,* December, 2023. P. 681-689.
11. Naimul Islam Suvon, Sadman Chowdhury Siam, Sadman Chowdhury Doctor of Philosophy admission prediction of Bangladeshi students into different classes of universities. *International Journal of Artificial Intelligence*. 2022. Vol. 11, No. 4. P. 1545-1553.
12. B. Wu, Z. Ke, M. Fu and Xia Y. SOUA: Towards Intelligent Recommendation for Applying for Overseas Universities. *International Conference on Intelligent Computing, Automation and Systems (ICICAS*), Chongqing, China, 2019. P. 124-128.
13. Nandal P. Deep Learning in diverse Computing and Network Applications Student Admission Predictor using Deep Learning. *Proceedings of the International Conference on Innovative Computing & Communications (ICICC),* March 28, 2020.
14. Navoneel C., Siddhartha C., Srinibas, R. A. Statistical Approach to Graduate Admissions' Chance Prediction. *Part of the Lecture Notes in Networks and Systems book series* (LNNS), Vol. 103. 2020. P. 333-340.
15. Acharya M. S., Armaan A. and Antony A. S. A Comparison of Regression Models for Prediction of Graduate Admissions. *International Conference on Computational Intelligence in Data Science (ICCIDS),* Chennai, India, 2019. P. 1-5.
16. Guabassi E., Bousalem I., Marah Z., Qazdar, A. A Recommender System for Predicting Students' Admission to a Graduate Program using Machine Learning Algorithms. *International Journal of Online and Biomedical Engineering (iJOE).* 2021. Vol. 17., No. 2. P. 135–147.
17. Callistus Obunadik Selection of best model to predict graduate admission into United States of American Universities using University of California, Los Angeles as a Case Study. *APSU Teaching Mathematics Conference*, Clarksville, Tennessee, USA, 2022.
18. Protikuzzaman Md., Mrinal Kanti Baowaly, Maloy Kumar Devnath and Bikash Chandra Singh. Predicting Undergraduate Admission: A Case Study in Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh. *International Journal of Advanced Computer Science and Applications (IJACSA).* 2020. Vol. 11, No. 12. P. 138-145.
19. Pavitha N., Ingale V., Verma V., Yeole A., Zawar S., Jamadar Z. Design Engineering Comparative analysis of Regression Algorithms for College Prediction. *Design Engineering Journal*. 2021.Vol.9. P. 6631-6643
20. Рейтингові списки вступників. URL: https://vstup2021.lpnu.ua (revised on 02.01.2023).
21. Ahmedbahaaaldin Ibrahem Ahmed Osman, Ali Najah Ahmed, Ming Fai Chow, Yuk Feng Huang, Ahmed El-Shafie, Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia, *Ain Shams Engineering Journal.* 2021. Vol. 12, No. 2. P.1545-1556.
22. Orange Visual Programming. URL: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/index.html (revised on 02.01.2023).

Khrystyna Zub **Христина Зуб**	Ph.D. student, Department of Social Communication and Information Activities, Lviv Polytechnic National University, Lviv, Ukraine. e-mail: khrystyna.v.zub@lpnu.ua https://orcid.org/0000-0001-6476-7305	аспірант кафедри соціальних комунікацій та інформаційної діяльності, Національний університет «Львівська політехніка», Львів, Україна
Pavlo Zhezhnych **Павло Жежнич**	DScTech., Prof., Department of Social Communication and Information Activities, Lviv Polytechnic National University, Lviv, Ukraine. e-mail: pavlo.i.zhezhnych@lpnu.ua https://orcid.org/0000-0002-2044-5408	д.т.н., проф., кафедри соціальних комунікацій та інформаційної діяльності, Національний університет «Львівська політехніка», Львів, Україна

UDC 004.7
https://doi.org/10.31891/csit-2023-1-12

Oleksandr MELNYCHENKO
Khmelnytskyi National University

METHOD OF REAL-TIME VIDEO STREAM SYNCHRONIZATION IN THE WORKING ENVIRONMENT OF AN APPLE ORCHARD

Monitoring and analyzing the state of harvest in an apple orchard is essential for efficient horticulture. Unmanned aerial vehicles (UAVs) have been increasingly used for this purpose due to their ability to capture high-resolution images and videos of the orchard from different perspectives. However, synchronizing the video streams from multiple UAVs in real-time presents a significant challenge. The traditional controller-worker architecture used for video stream synchronization is prone to latency issues, which can negatively impact the accuracy of the monitoring system. To address this issue, the authors propose a decentralized method using a consensus algorithm that allows the group of UAVs to synchronize their video streams in real time without relying on a centralized controller device. The proposed method also addresses the challenges of limited network connectivity and environmental factors, such as wind and sunlight. The automated system that utilizes the proposed method was tested in an actual apple orchard. The experimental results show that the proposed approach achieves real-time video stream synchronization with minimal latency and high accuracy. As such, the SSIM index varies from 0.79 to 0.92, with an average value of 0.87, and the PSNR index – varies from 22 to 39, which indicates the decent quality of the received information from combined images. Meanwhile, the effectiveness of the developed system with the proposed approach was proven, which is confirmed by a high average value of 82.69% of the reliability indicator of detecting and calculating the number of fruit fruits and a low average level of type I (14.67%) and II (18.33%) errors. Overall, the proposed method provides a more reliable and efficient approach to real-time video stream synchronization in an apple orchard, which can significantly improve the monitoring and management of apple orchards.

Keywords: real-time video stream, synchronization, image stitching, apple orchard, unmanned aerial vehicles, decentralized approach.

Олександр МЕЛЬНИЧЕНКО
Хмельницький національний університет

МЕТОД СИНХРОНІЗАЦІЇ ВІДЕОПОТОКІВ В РЕЖИМІ РЕАЛЬНОГО ЧАСУ В РОБОЧОМУ СЕРЕДОВИЩІ ЯБЛУНЕВОГО САДУ

Моніторинг та аналіз стану врожайності в яблуневому саду є важливими для здійснення ефективного садівництва. Безпілотні літальні апарати (БПЛА) усе частіше використовуються для цієї мети завдяки їхній здатності знімати зображення та відео високої роздільної здатності саду з різних ракурсів. Однак синхронізація відеопотоків із кількох БПЛА в реальному часі може спричиняти низку технічних проблем. Так, традиційна архітектура управління групою БПЛА під назвою «контролер-працівник», яка використовується для синхронізації відеопотоку, схильна до проблем із затримкою, що може негативно вплинути на точність системи моніторингу. Тому, для вирішення подібної проблеми, у цій роботі пропонується децентралізований метод із використанням консенсусного алгоритму, який дає змогу групі БПЛА синхронізувати свої відеопотоки в режимі реального часу без огляду на централізований пристрій керування. Запропонований метод також вирішує проблеми обмеженого підключення до мережі та враховує негативний вплив чинників навколишнього середовища, таких як пориви вітру та високу хмарність. Розроблена автоматизована система, що ґрунтується на запропонованому методі, може працювати в середовищах із низьким рівнем підключення та справлятися з проблемами, пов'язаними із чинниками робочого середовища фруктового саду. У результаті проведення експериментальних досліджень над автоматизованою системою встановлено, що запропонований підхід забезпечує синхронізацію відеопотоку в реальному часі з мінімальною затримкою та високою точністю. Зокрема, оцінка синхронізації відеопотоків за індексом SSIM коливається від 0,79 до 0,92 із середнім значенням 0,87, а за індексом PSNR – від 22 до 39, що свідчить про високу ефективність роботи розробленої системи з відеопотоками та високою якістю отриманої інформації з комбінованих зображень. Заразом було доведено ефективність розробленої системи із запропонованим підходом, що підтверджується високим середнім значенням 82,69 % показника достовірності виявлення яблук та низьким середнім рівнем похибок I (14,67 %) та II (18,33 %) роду. Загалом запропонований метод забезпечує більш надійний та ефективний підхід до синхронізації відеопотоку в реальному часі в яблуневому саду, що може значно покращити моніторинг та управління яблуневими садами.

Ключові слова: відеопотік у реальному часі, синхронізація, об'єднання зображень, яблуневий сад, безпілотні літальні апарати, децентралізований підхід.

Introduction

Apple orchards often face numerous challenges, including pest infestations, weather changes, disease outbreaks, and labor shortages. These challenges can lead to reduced crop yields, increased costs, and even crop failure [1]. Furthermore, traditional methods of monitoring and managing orchards [2], such as manual inspection, can be time-consuming, labor-intensive, and often yield incomplete or inaccurate information.

To address these challenges, there is a growing need for implementing information technologies in apple orchards. Using technologies such as drones, sensors, and computer vision can provide real-time data on crop health, soil moisture, temperature, and other factors impacting fruit growth and yield [3, 4]. This data can be used to optimize fruit management strategies, such as targeted irrigation, pest control, and automatic detection and calculation of the amount of harvest that may increase fruit yields and reduce costs.

Specifically, monitoring and analyzing the growth and condition of apples in a fruit orchard is essential for effective orchard management. Unmanned aerial vehicles (UAVs) have been increasingly used for this purpose due

to their ability to capture high-resolution images and videos of the orchard from different perspectives [4]. However, synchronizing the video streams from multiple UAVs in real-time presents a significant challenge [5].

The problem of real-time video stream synchronization from UAVs in the working environment of an apple orchard is the focus of many recent kinds of research. The main challenge arises from UAVs being in constant motion and subject to various environmental factors, such as wind and sunlight [5, 6]. Weather conditions cause significant variations in the captured video streams, making it difficult to align them accurately in real-time.

Related works

Over the past decade, researchers have proposed several methods for real-time video stream synchronization from UAVs in the working environment of an apple orchard. One approach uses a controller-worker architecture [7], where one UAV is designated as the controller, and the other UAVs are designated as workers [8]. The controller UAV generates a synchronization signal transmitted to the worker UAVs through a wireless connection [9]. The worker UAVs then adjust their internal clocks to match the controller UAV, ensuring that all UAVs are synchronized.

Another approach [10] is to use visual odometry, which is a technique for estimating the motion of a vehicle by analyzing the changes in the images captured by its camera. In other work [11], each UAV is equipped with a visual odometry system that estimates its motion in real time. The estimated motion is then used to align the UAV video streams. One of the challenges in using visual odometry is the accuracy of the motion estimation, which can be affected by various factors such as camera calibration, image noise, and scene complexity. To address this challenge, researchers have proposed various techniques for improving the accuracy of visual odometry, such as using multiple sensors [12] and incorporating deep learning algorithms [13].

Some researchers have also proposed using sensor fusion to synchronize the video streams from multiple UAVs, like in [14]. In another study [15], each UAV has multiple sensors, such as GPS, inertial measurement units, and magnetometers. Such approaches employ advanced algorithms to fuse the sensor data to estimate the position and orientation of each UAV in real time, yet commonly with low accuracy. In this case, the estimated position and orientation are then used to align the video streams captured by the UAVs.

Overall, the problem of real-time video stream synchronization from UAVs in the working environment of an apple orchard is a challenging research problem with significant implications for orchard management. Accurately synchronizing the video streams from multiple UAVs can enable more accurate tracking of the growth and condition of apple trees [16], leading to improved crop yield optimization. Additionally, the proposed methods for real-time video stream synchronization from UAVs have applications beyond apple orchards and can be applied in other dynamic environments where real-time video stream synchronization is essential.

Problem statement

From the literature review, it was observed that traditional approaches to video stream synchronization rely on using a centralized controller-worker architecture [17], where one device acts as a controller, and the other devices act as workers. However, this architecture is prone to latency issues, which can result in delays in video stream synchronization and negatively impact the accuracy of the monitoring system. Additionally, the working environment of an apple orchard presents additional challenges, such as limited network connectivity and environmental factors, such as wind and sunlight, which can further exacerbate latency issues and make real-time video stream synchronization more challenging.

Therefore, there is a need for a new method of real-time video stream synchronization in the working environment of an apple orchard that can effectively address the challenges posed by multiple UAVs and environmental factors. This method should be reliable, efficient, and able to synchronize video streams in real-time with minimal latency while also being able to operate in low-connectivity environments and handle the challenges posed by environmental factors. Such a method would significantly improve the monitoring and management of apple orchards, allowing for more accurate and efficient decision-making.

Method of real-time video stream synchronization

The automated multi-level system proposed for detecting and calculating the number of similar structural objects by a group of UAVs utilizes multiple hardware devices to capture video sequences of the target objects. This unique feature enables the system to efficiently process and analyze a large number of video streams in real time. The proposed system can be used in various applications, such as monitoring agricultural fields, detecting structural damage in buildings, and assessing disaster areas.

Synchronizing video streams from multiple UAVs can be complicated due to various factors, such as the type of video cameras used, speed differences in receiving video streams, and distorted or missing video streams. Differences in flight characteristics and video capture methods between UAVs from different manufacturers can also negatively affect the detection quality and accuracy of structural object calculations. To address these problems, this study proposes a new method for real-time video stream synchronization. The method involves merging video sequences obtained from each drone in a group during an operational mission into a single image of a fruit tree to prevent issues with receiving video sequences. The proposed method is implemented through several software blocks

combined into a single information system to create a behavioral signature, detect, and calculate the number of structural objects representing apples on trees. The method is depicted in fig. 1.

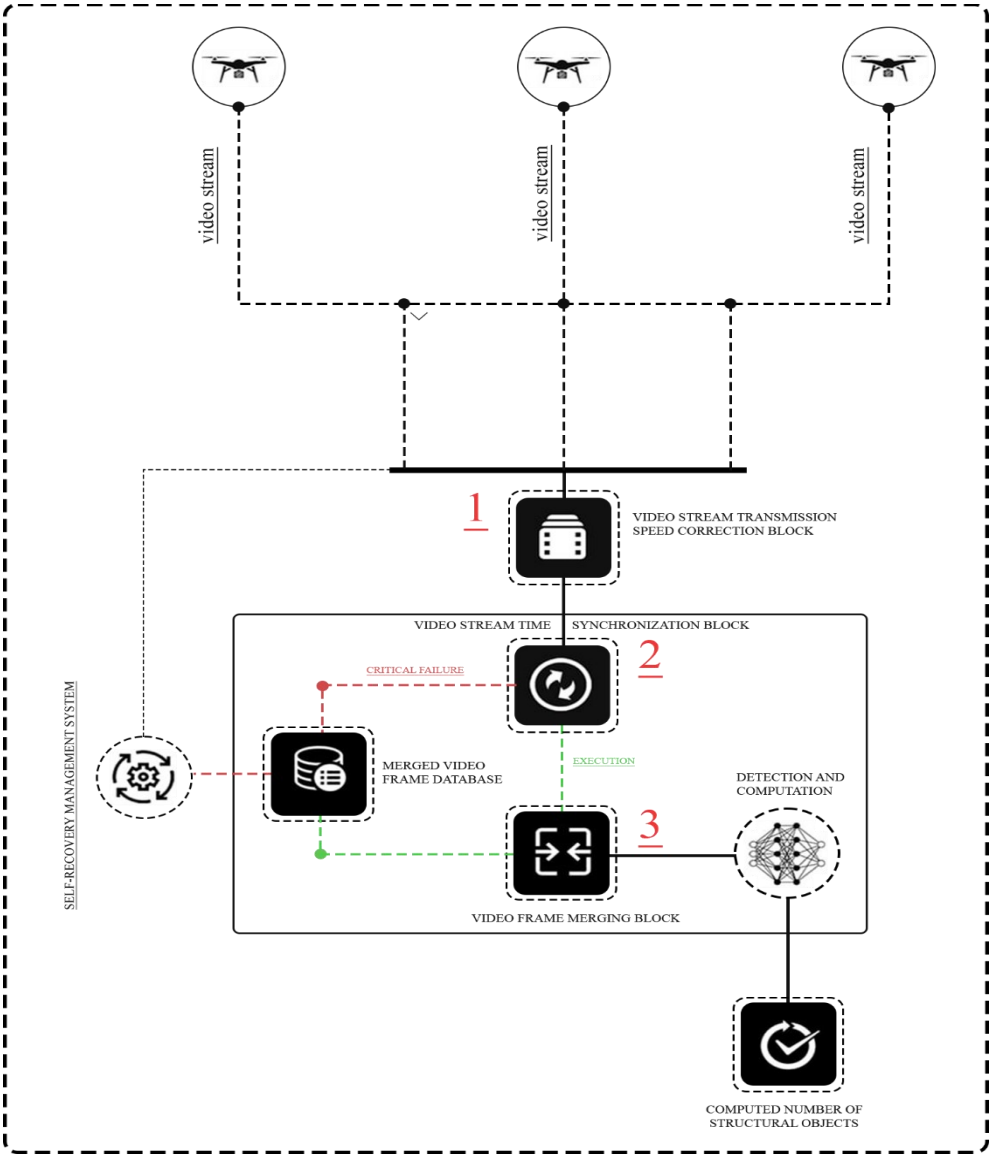


Fig. 1. The scheme of the method of real-time synchronization of video streams

The method consists of the following main components:

1. ***Block 1***: unit for adjusting the speed of transmission of video streams. While conducting a software mission in the working area, the group of UAVs generates video sequences that are transmitted over the network to the detection software module. However, these sequences enter the module at different times, which can be due to several factors, such as fluctuations in the network connection, the type of cameras used by the UAVs, and the speed at which the drones move in the working area, among others.

Given the characteristics of the video stream flow in the work environment, the speed adjustment unit constructs a software framework that incorporates mechanisms to process the transmission speed of video streams and establishes the functionality of halting and receiving a video stream. When the block for adjusting the transmission speed of a video stream transitions to the «waiting» state, it guarantees the reception of all video streams from each UAV simultaneously.

The block's generated software structure includes the following features [18]: 1) an identifier specific to the drone, 2) bytes set of the video frame, 3) the video sequence's frame rate, 4) the time of the speed correction block receiving the video frame from the drone, and 5) coding format of the video frame. When block completes its task, it transitions to the «execution» state and delivers the generated multiple program structures to the next execution block.

2. ***Block 2***: synchronization of video streams in time. The functioning of this block is based on the quantity of program structure sets received from the preceding block responsible for the speed correction of video streams. Initially, the video stream synchronization block inspects the number of program structure sets and processes the following scenarios accordingly:

2.1. A critical state is triggered if the number of program structure sets received by the synchronization block does not match the number of drones. The block then generates a request to retrieve the most recently saved program structure from the database of merged video frames. If the database contains the requested program structure, it is sent to the self-recovery module upon request. However, if the requested program structure is not in the database, a critical request is immediately issued to terminate the operation since it suggests that the UAV group failed to complete its mission. First, the video frame fusion unit checks the equivalence of the generation time of a set of program data structures. Performing a characteristic time check ensures the integrity of the generated data while creating one program data structure.

2.2. If video streams are successfully synchronized, the behavioral signature transitions to the "video sequence storyboarding" state. However, each drone may have different hardware that encodes video frames differently. Therefore, to ensure the unit can detect and calculate structural objects, the frames are converted into a single software format as an image. A derivative unit is developed that decodes the video frames into the required system format and creates program structures. The decoding time is recorded and added to the structure. As a result, the output of this block is a set of program structures that act as input data for the next block, which merges the video frames.

3. ***Block 3***: Unit for merging video frames. To ensure that the software objects used by the neural network have accurate geometric parameters, video frames captured by drones at various heights and angles are transformed using algebraic image transformation algorithms. The video cameras used by drones from different manufacturers have varying capture widths, and the weather conditions in the working environment can affect the camera's stability and distort the visual area. The dynamic nature of the working environment means that weather conditions and wind gusts can change during the UAV group's mission. Fig. 2 shows a diagram of the execution process of the video frame fusion block.

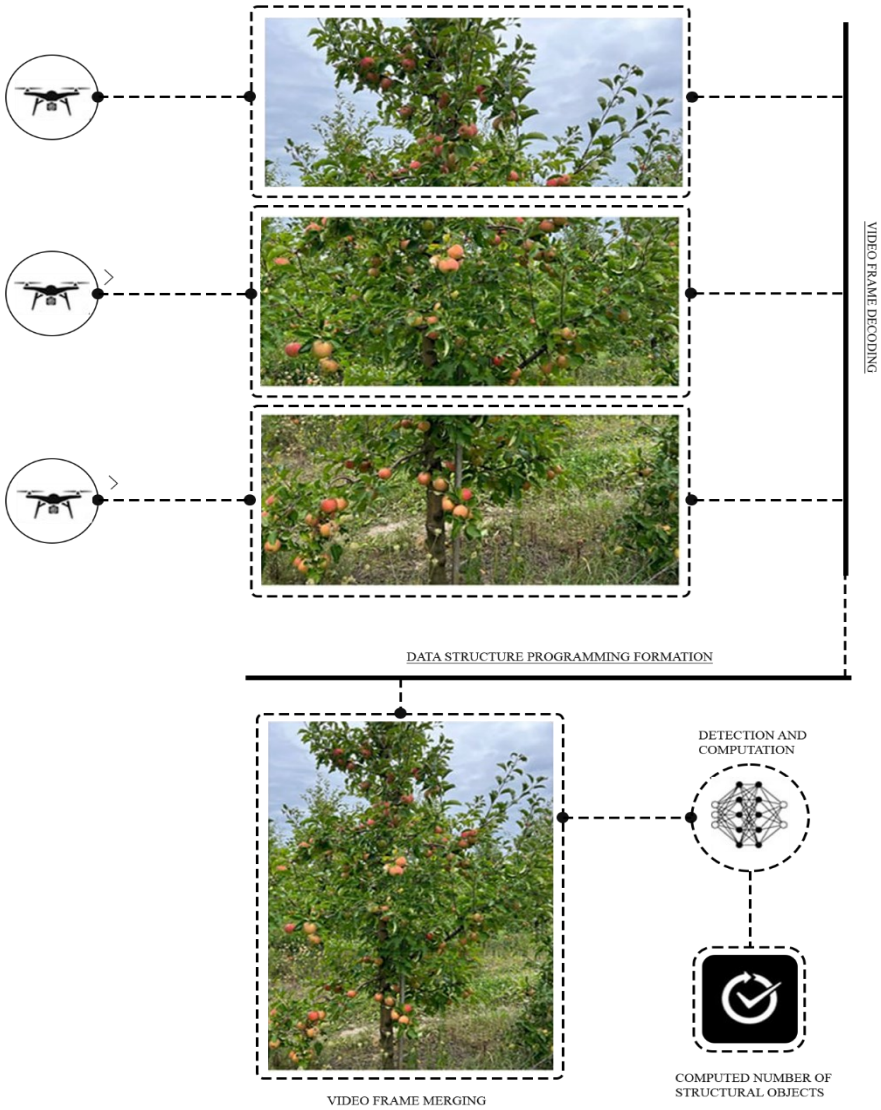


Fig. 2. Process diagram of the video frame fusion unit

The video frames undergo further transformation through the following steps:

Step 1: The interlaced video frame represented by a set of bytes is converted into a file. The software mechanisms then check the file's metadata, including file type, geometry, color space, resolution, and channel depth, by determining the file's signatures.

1.1) The set of images obtained in Step 1 is then corrected for rotation angle using an affine transformation, resulting in transformed video frames with coinciding geometric feature values.

1.2) The converted images are then sent to the software merge function.

Step 2: The corrected images are merged into a single image by fusing them.

2.1) The merged image may contain graphical artifacts, such as overexposed or underexposed areas and varying depth of field between frames. To address this issue, a software filter defined in the system is applied to mask the transitions between the frames. This filtering ensures that the merged video frames appear as a single image without noticeable transitions and with transparent edges.

2.2) The software compression engine receives the filtered video frames.

2.3) The video frame compression process involves an affine transformation that relies on the geometric transformation parameters of the neural network algorithms' input data. Since the video frames captured by the UAVs are rectangular, the compression mechanism converts them to a square shape for maximum efficiency in detecting and calculating the number of structural objects. This step results in a set of video frames ready for merging.

2.4) The set of prepared video frames from step 2.3) is then passed to the software fusion function.

2.5. Merging multiple video frames into one complete image produces a matrix-type object program data structure. Each element of the matrix corresponds to the color code value of a single graphic pixel. This matrix forms a continuous representation of a fruit tree, where all the single video frames from different drones combine into one image.

2.6) The data obtained in step 2.5) is then stored in the internal database of merged video frames.

2.7) The matrix software data structure is sent to the software module responsible for detecting and calculating the number of structural objects with similar characteristics.

The successful execution of the video frame fusion block results in a program data structure represented by a matrix of color codes. The fusion unit incorporates a functionality element that stores merged video frames in the internal database, which ensures system integrity in the event of a critical failure. The video frame fusion unit uses image conversion mechanisms to automatically process all video streams received during the UAV group's program mission. The system can identify critical failures that distort the data structure's integrity and store them in an error log to prevent their use as input parameters for further processing. Therefore, the video stream synchronization component ensures data integrity and prevents the system from processing distorted information.

Experimental results

The structural similarity index (SSIM) and the peak visual signal-to-noise ratio (PSNR) index [19] were used to evaluate the effectiveness of the real-time video stream synchronization method that is proposed in this work.

The SSIM index is formalized as follows

$$\mathrm{SSIM} = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}, \tag{1}$$

where μ_x and μ_y are the mean values of pixels in the input and merged images, respectively, σ_x^2 and $\sigma_н^2$ are the standard deviations of pixels in the input and merged images, respectively, σ_{xy} is the covariance between pixels in both images, C_1 and C_2 are constant values that allow stabilizing the resulting value of the formula.

The PSNR index is formalized by the formula

$$PSNR = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right), \tag{2}$$

where MAX_I is the maximum pixel value in the original image *I*;

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[I\left(i,j\right) - K\left(i,j\right)\right]^2,$$

where *I* is a set of three original images of total $m \times n$ pixels, *K* is the $m \times n$ merged images.

Achieving the efficiency of video stream synchronization according to (1) consists of obtaining a value from 0.5 to 1, considered a high-efficiency value; at the same time, an SSIM value in the range of 0 to 0.49 indicates

ineffective synchronization. Formula (2) represents the degree of quality of the image obtained because of the merging operation; the value of the PSNR index is calculated as the ratio between the maximum possible power of the visual signal and the noise present in the image; the higher the value, the better the quality of the received image.

Table 1 shows the results of the video stream synchronization module of the automated system implementing the corresponding method for a stream of 12 consecutive groups of video frames randomly selected for testing; each group contains three video frames obtained from three UAVs, which are further combined into one image.

Table 1

The study outcomes of the efficiency of the video stream synchronization method achieved by the UAV group.

Index of a merged image	SSIM	PSNR	Index of a merged image	SSIM	PSNR
1	0.90	35.20	7	0.86	31.43
2	0.45	27.22	8	0.47	27.87
3	0.72	37.50	9	0.90	30.12
4	0.85	29.11	10	0.83	31.54
5	0.87	36.90	11	0.52	28.91
6	0.91	39.10	12	0.86	30.36

From table 1, SSIM performance ranges from 0.79 to 0.92, with an average value of 0.87. At the same time, images for which the value of the SSIM index is less than 0.50 are considered distorted by the system. Currently, those merged images with a PSNR index value greater than 30 are considered high quality; at the same time, PSNR values less than 30 indicate low image quality, which may be caused by external factors of the working environment (strong gusts of wind, precipitation, etc.).

If the current image has an SSIM index value less than 0.50, and the SSIM index value is less than 30, then this image will be considered distorted by the system and will therefore be discarded and will not go to the next module of detecting and calculating the number of structural objects.

In order to evaluate the practical validity of the proposed method, we applied an object detector [20] to the merged images to identify and classify apples that appear on the trees. A comparison of the estimates of the types I and II errors obtained by the UAV group during experiments under different weather conditions is shown in fig. 3.

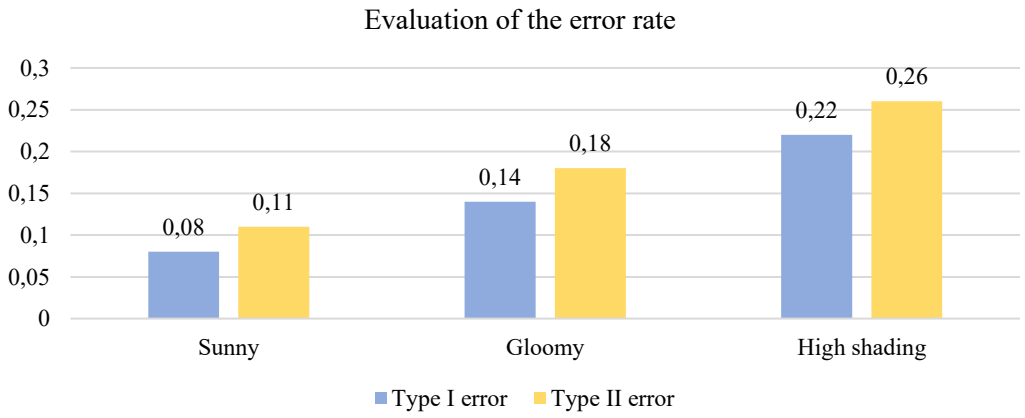


Fig. 3. Comparing the evaluation of errors made by the UAV group in various weather conditions

Fig. 3 demonstrates that the quality of fruit recognition in natural conditions is heavily influenced by weather factors and the presence of shading caused by other trees in the target work zones. These factors, in combination with visual noise such as leaves and tree branches, make it challenging to identify and track target objects in real-time dynamics. This is mainly due to the limited coverage angle of UAV cameras.

The conducted experiments confirmed the effectiveness of using a group of UAVs as part of an automated system for flying over an orchard in various weather conditions. The results showed that under sunny and overcast conditions with soft shade, the system had high accuracy and reliability with type I and II errors at 8% and 11%, respectively; under sunny conditions, and 14% and 18%, respectively; under cloudy weather. However, the errors were higher at 22% and 26% under highly shaded conditions, respectively. It is important to note that the presence of visual noise in the orchard, such as fruits being covered by leaves and branches, means that the UAV group and the automated system cannot achieve 100% efficiency in natural conditions, which is a promising area for further research.

An experimental study was conducted to evaluate the effectiveness of the developed automated system in detecting and calculating the number of fruits in natural conditions. The evaluation criteria included: a) the E index to determine the effectiveness of automatic route determination for the UAV group, b) the accuracy indicators for fruit detection and type I and II errors, and c) the indicators for the effectiveness of real-time synchronization of video

frames SSIM and PSNR. The results of the experiments showed that the developed automated system is efficient, as evidenced by a high-reliability indicator of 82.69% on average for detecting and calculating the number of fruits.

Conclusions

The developed automated system with the proposed method can detect and count apples in real time in an orchard. Specifically, the system can receive multiple video frames in real-time from several UAV cameras, synchronize these video frames with each other into one informational data structure, and optimize image quality to improve the detection of apples. The system's video stream synchronization is evaluated based on the SSIM index, which ranges from 0.79 to 0.92 with an average value of 0.87, and the PSNR index, which ranges from 22 to 39. These results indicate the system's high efficiency with video streams and the decent quality of the information received from combined images. Moreover, the effectiveness of the developed automated system was confirmed by a high average value of 82.69% of the reliability indicator of detecting and calculating the number of fruit fruits and a low average level of type I (14.67%) and II (18.33%) errors.

Further research will be conducted to explore the potential of integrating deep learning algorithms into the system to improve the accuracy and efficiency of image processing.

References

1. Lauri P.-É., Simon S. Advances and challenges in sustainable apple cultivation. *Achieving sustainable cultivation of temperate zone tree fruits and berries*. Burleigh Dodds Science Publishing, 2019. 28 p.

2. Segarra J., Buchaillot M. L., Araus J. L. et al. Remote sensing for precision agriculture: Sentinel-2 improved features and applications. *Agronomy*. 2020. Vol. 10, No. 5. P. 641. DOI: https://doi.org/10.3390/agronomy10050641

3. Paul K., Chatterjee S. S., Pai P. et al. Viable smart sensors and their application in data driven agriculture. *Computers and Electronics in Agriculture*. 2022. Vol. 198. P. 107096. DOI: https://doi.org/10.1016/j.compag.2022.107096

4. Zhang C., Valente J., Kooistra L. et al. Orchard management with small unmanned aerial vehicles: A survey of sensing and analysis approaches. *Precision Agriculture*. 2021. Vol. 22, No. 6. P. 2007–2052. DOI: https://doi.org/10.1007/s11119-021-09813-y

5. Zhou H., Hu F., Juras M. et al. Real-time video streaming and control of cellular-connected UAV system: Prototype and performance evaluation. *IEEE Wireless Communications Letters*. 2021. Vol. 10, No. 8. P. 1657–1661. DOI: https://doi.org/10.1109/LWC.2021.3076415

6. Koutalakis P., Tzoraki O., Gkiatas G. et al. Using UAV to capture and record torrent bed and banks, flood debris, and riparian areas. *Drones*. 2020. Vol. 4, No. 4. P. 77. DOI: https://doi.org/10.3390/drones4040077

7. Pirihan E. Near real-time web-based mapping of live video streams from unmanned aerial vehicles : Thesis. Ankara, Turkey : Hacettepe Üniversitesi, 2022. 132 p.

8. Mao W., Liu H., Hao W. et al. Development of a combined orchard harvesting robot navigation system. *Remote Sensing*. 2022. Vol. 14, No. 3. P. 675. DOI: https://doi.org/10.3390/rs14030675

9. Lim W. Y. B., Garg S., Xiong Z. et al. UAV-assisted communication efficient federated learning in the era of the artificial intelligence of things. *IEEE Network*. 2021. Vol. 35, No. 5. P. 188–195. DOI: https://doi.org/10.1109/MNET.002.2000334

10. Duan R., Paudel D. P., Fu C. et al. Stereo orientation prior for UAV robust and accurate visual odometry. *IEEE/ASME Transactions on Mechatronics*. 2022. Vol. 27, No. 5. P. 3440–3450. DOI: https://doi.org/10.1109/TMECH.2022.3140923

11. McConville A., Bose L., Clarke R. et al. Visual odometry using pixel processor arrays for unmanned aerial systems in GPS denied environments. *Frontiers in Robotics and AI*. 2020. Vol. 7. P. 126. DOI: https://doi.org/10.3389/frobt.2020.00126

12. Kaygusuz N., Mendez O., Bowden R. AFT-VO: Asynchronous fusion transformers for multi-view visual odometry estimation. *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2022)* : Proceedings. (Kyoto, Japan, October 23-27, 2022), IEEE Inc., 2022. P. 2402–2408. DOI: https://doi.org/10.1109/IROS47612.2022.9981835. P. 2402–2408

13. Radiuk P., Pavlova O., Avsiyevych V., Kovalenko V. Convolutional neural network for parking slots detection. *The 3rd International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS-2022)* : CEUR-Workshop Proceedings. Vol. 3156. (Khmelnytskyi, 23-25 March 2022). Khmelnytskyi, 2022. P. 284–293. URL: http://ceur-ws.org/Vol-3156/paper21.pdf

14. Gupta A., Fernando X. Simultaneous localization and mapping (SLAM) and data fusion in unmanned aerial vehicles: Recent advances and challenges. *Drones*. 2022. Vol. 6, No. 4. P. 85. DOI: https://doi.org/10.3390/drones6040085

15. Du H., Wang W., Xu C. et al. Real-time onboard 3D state estimation of an unmanned aerial vehicle in multi-environments using multi-sensor data fusion. *Sensors*. 2020. Vol. 20, No. 3. P. 919. DOI: https://doi.org/10.3390/s20030919

16. Mademlis I., Torres-González A., Capitán J. et al. A multiple-UAV architecture for autonomous media production. *Multimedia Tools and Applications*. 2023. Vol. 82, No. 2. P. 1905–1934. DOI: https://doi.org/10.1007/s11042-022-13319-8

17. Rech L. C., Junior L. B., Berger G. S. et al. Proposal of a visual positioning architecture for master-slave autonomous UAV applications. *ROBOT2022: Fifth Iberian Robotics Conference (ROBOT-2022)* : Proceedings. Vol. 590. Springer, Cham : Springer International Publishing, 2023. P. 365–375. DOI: https://doi.org/10.1007/978-3-031-21062-4_30. C. 365–375

18. Savenko O., Lysenko S., Nicheporuk A. et al. Metamorphic viruses' detection technique based on the equivalent functional block search. *13th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge* : CEUR-Workshop Proceedings. Vol. 1844. (Kyiv, 15-18 May 2017). Kyiv, 2017. P. 555–568. URL: https://ceur-ws.org/Vol-1844/10000555.pdf

19. Sara U., Akter M., Uddin M. S. Image quality assessment through FSIM, SSIM, MSE and PSNR—A comparative study. *Journal of Computer and Communications*. 2019. Vol. 7, No. 3. P. 8–18. DOI: https://doi.org/10.4236/jcc.2019.73002

20. Barmak O., Radiuk P. Web-based information technology for classifying and interpreting early pneumonia based on fine-tuned convolutional neural network. *Computer systems and information technologies*. 2021. Vol. 3, No 1. P. 12–18. DOI: https://doi.org/10.31891/CSIT-2021-3-2

Oleksandr Melnychenko **Олександр Мельниченко**	PhD student of the Department of Computer Engineering and Information Systems, Khmelnytskyi National University, Khmelnytskyi, Ukraine. e-mail: oleksandr.melnychenko@live.com https://orcid.org/0000-0001-8565-7092	аспірант кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет, Хмельницький, Україна.

Full requirements for the design of the manuscript
Повні вимоги до оформлення рукопису
http://csitjournal.khmnu.edu.ua/

No editorial responsibility is required for the content of messages sub.
За зміст повідомлень редакція відповідальності не несе

To print 30.03.2023. Mind. Printing. Arch. 10,35. Obl.-vid. Arch. 9,72
Format 30x42 / 4, offset paper. Another risography.
Overlay 100, deputy. №

Підп. до друку 30.03.2023. Ум. друк. арк. 10,35. Обл.-вид. арк. 9.72
Формат 30х42/4, папір офсетний. Друк різографією.
Наклад 100, зам. №

Replication is made from the original layout, made edited
by the magazine "Computer Systems and Information Technology"

Тиражування здійснено з оригінал-макету, виготовленого
редакцією журналу “Комп’ютерні системи та інформаційні технології”

Editorial and publishing center of Khmelnytskyi national university
29016, Khmelnytskyi, street Institutska, 7/1, tel. (0382) 72-83-63

Редакційно-видавничий центр Хмельницького національного університету
29016, м. Хмельницький, вул. Інститутська, 7/1, тел. (0382) 72-83-63