Informatics Control Measurement
in Economy and Environment Protection

4/2023

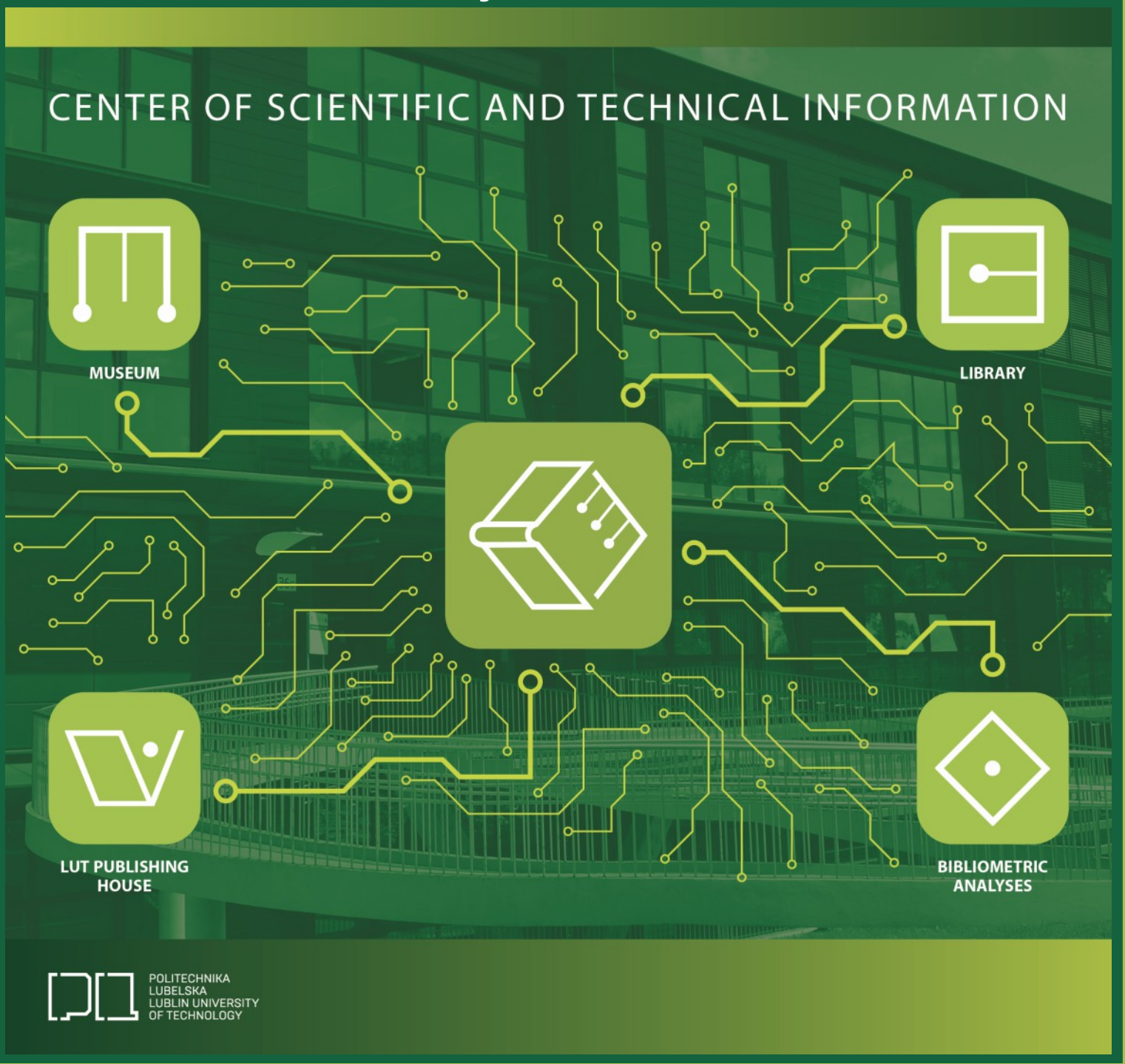
Informatyka Automatyka Pomiary

w gospodarce i ochronie środowiska

www.e-IAPGOS.pl

ISSN 2083-0157

Kwartalnik Naukowo-Techniczny



4/2023

październik – grudzień

Wydanie pod redakcją naukową
prof. dr hab. inż. Waldemara Wójcika

INFORMATYKA AUTOMATYKA POMIARY
W GOSPODARCE i OCHRONIE ŚRODOWISKA
Informatics Control Measurement in Economy and Environment Protection

p-ISSN 2083-0157, e-ISSN 2391-6761, www.e-iapgos.pl

PRINTING HOUSE – DRUKARNIA

PPH Remigraf Sp. z o.o.
ul. Dźwigowa 61, 01-376 Warszawa
https://remigraf.pl/

nakład: 100 egzemplarzy

OTHER INFORMATION – INNE INFORMACJE

Czasopismo jest indeksowane w bazach:

DOAJ	doaj.org
BazTech	baztech.icm.edu.pl
IC Journals Master List	www.journals.indexcopernicus.com
Google Scholar	scholar.google.pl
POL-index	pbn.nauka.gov.pl
Sherpa RoMEO	www.sherpa.ac.uk
OAJI	oaji.net
SCOPUS	www.scopus.com
EBSCO	www.ebsco.com

Czasopismo *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska* zostało objęte finansowaniem przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach programu *Wsparcie dla czasopism naukowych* w latach 2019-2020.

Czasopismo znajduje się w wykazie czasopism naukowych opublikowanym w Komunikacie Ministra Edukacji i Nauki z dnia 17 lipca 2023 r. z późn. zm., Unikatowy Identyfikator Czasopisma: 200167 – z przypisaną liczbą punktów przyznawanych za publikację artykułu równą 20.

Zasady publikowania artykułów, przygotowania tekstów, zasady etyczne, procedura recenzowania, wykazy recenzentów oraz pełne teksty artykułów dostępne są na stronie internetowej czasopisma:

www.e-iapgos.pl

W celu zwiększenia oddziaływania czasopisma w środowisku naukowym redakcja zaleca:

- w artykułach publikowanych w IAPGOS cytować artykuły z renomowanych czasopism międzynarodowych (szczególnie indeksowanych w bazach Web of Science oraz Scopus) używając oficjalnych skrótów nazw czasopism,
- w artykułach publikowanych w innych czasopismach (zwłaszcza indeksowanych w bazach Web of Science oraz Scopus) cytować prace publikowane w IAPGOS – zwłaszcza posługując się numerami DOI, np.:
Kluszczyński K. *Modelowanie – umiejętność czy sztuka?* Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska – IAPGOS, 1/2016, 4–15, https://doi.org/10.5604/20830157.1193833.

CONTENTS – SPIS TREŚCI

1. **Valerii Kryvonosov, Oleg Avrunin, Serhii Sander, Volodymyr Pavlov, Liliia Martyniuk, Bagashar Zhumazhanov**
A usage of the impedance method for detecting circulatory disorders to determine the degree of limb ischemia
Impedancyjna metoda wykrywania zaburzeń krążenia krwi do określenia stopnia niedokrwienia kończyny 5

2. **Konrad Witkowski, Mikołaj Wieczorek**
Usage of artificial neural networks in the diagnosis of knee joint disorders
Zastosowanie sztucznych sieci neuronowych w diagnozie schorzeń stawu kolanowego 11

3. **Oumaima Majdoubi, Achraf Benba, Ahmed Hammouch**
Comprehensive machine learning and deep learning approaches for Parkinson's disease classification and severity assessment
Kompleksowe metody uczenia maszynowego i uczenia głębokiego do klasyfikacji choroby Parkinsona i oceny jej nasilenia 15

4. **Mamun Ahmed, Salma Binta Islam, Aftab Uddin Alif, Mirajul Islam, Sabrina Motin Saima**
AI empowered diagnosis of pemphigus: a machine learning approach for automated skin lesion detection
Diagnostyka pęcherzycy z wykorzystaniem sztucznej inteligencji: podejście oparte na uczeniu maszynowym do automatycznego wykrywania zmian skórnych 21

5. **Mohamed Bal-Ghaoui, My Hachem El Yousfi Alaoui, Abdelilah Jilbab, Abdennaser Bourouhou**
Optimizing ultrasound image classification through transfer learning: fine-tuning strategies and classifier impact on pre-trained inner-layers
Optymalizacja klasyfikacji obrazów ultrasonograficznych techniką transfer learning: strategie dostrajania i wpływ klasyfikatora na wstępnie wytrenowane warstwy wewnętrzne 27

6. **Swarajya Madhuri Rayavarapu, Tammineni Shanmukha Prasanthi, Gottapu Santosh Kumar, Gottapu Sasibhushana Rao, Gottapu Prashanti**
A generative model for Deep Fake Augmentation of phonocardiogram and electrocardiogram signals using LSGAN and Cycle GAN
Generatywny model z Deep Fake Augumentation dla sygnałów z fonokardiogramu oraz elektrokardiogramu w strukturach LSGAN oraz Cycle GAN 34

7. **Kamaran H. Manguri, Aree A. Mohammed**
Smart optimizer selection technique: a comparative study of modified DensNet201 with other deep learning models
Inteligentna technika wyboru optymalizatora: badanie porównawcze zmodyfikowanego modelu DensNet201 z innymi modelami głębokiego uczenia 39

8. **Yuliia Kholodniak, Yevhen Havrylenko, Serhii Halko, Volodymyr Hnatushenko, Olena Suprun, Tatiana Volina, Oleksandr Miroshnyk, Taras Shchur**
Improvement of the algorithm for setting the characteristics of interpolation monotone curve
Ulepszenie algorytmu wyznaczanie charakterystyki interpolacyjnej krzywej monotonicznej 44

9. **Marcin Cieśla, Mariusz Dzieńkowski**
An analysis of the implementation of accessibility tools on websites
Analiza implementacji narzędzi dostępności na stronach WWW 51

10. **Denys Ratov, Oleh Zakhozhai**
Interaction method between WebView objects in hybrid Java applications
Metoda interakcji pomiędzy obiektami WebView w hybrydowych aplikacjach Java 57

11. **Szymon Binek, Jakub Góral**
BrowserSpot – a multifunctional tool for testing the front-end of websites and web applications
BrowserSpot – multifunkcjonalne narzędzie do testowania front-endu stron internetowych oraz aplikacji sieciowych 61

12. **Roman Kvyetnyy, Yuriy Bunyak, Olga Sofina, Oleksandr Kaduk, Orken Mamyrbayev, Vladyslav Baklaiev, Yeraliyeva Bakhyt**
Advertising bidding optimization by targeting based on self-learning database
Optymalizacja ofert reklamowych poprzez ukierunkowanie w oparciu o samouczącą się bazę danych 66

13. **Dominik Choma, Kinga Chwaleba, Mariusz Dzieńkowski**
The efficiency and reliability of backend technologies: Express, Django, and Spring Boot
Wydajność i niezawodność technologii wytwarzania aplikacji internetowych strony serwera: Express, Django oraz Spring Boot 73

14. **Artem Yurchenko, Anzhela Rozumenko, Anatolii Rozumenko, Roman Momot, Olena Semenikhina**
Cloud technologies in education: the bibliographic review
Technologie chmurowe w edukacji: przegląd bibliograficzny 79

15. **Zakaria A. Hamed Alnaish, Safwan O. Hasoon**
Hybrid binary whale optimization algorithm based on taper shaped transfer function for software defect prediction
Hybrydowy, binarny algorytm WOA oparty na transmitancji stożkowej do prognozowania defektów oprogramowania 85

16. **Andrzej Szymon Borkowski, Jakub Brożyna, Joanna Litwin, Weronika Rączka, Aleksandra Szponarowicz**
Use of the CDE environment in team collaboration in BIM
Wykorzystanie platformy CDE we współpracy zespołowej w BIM 93

17. **Dmytro Yevgrafov, Yurii Yaremchuk**
Asymptotically optimal algorithm for processing side radiation signals from monitor screens on liquid crystal structures
Asymptotycznie optymalny algorytm przetwarzania sygnałów promieniowania bocznego z ekranów monitorów LCD 99

18. **Mariusz Ostrowski**
AC power regulation techniques for renewable energy sources
Techniki regulacji mocy odbiorników AC dla odnawialnych źródeł energii 103

19. **Mykhailo Burbelo, Oleksii Babenko, Yurii Loboda, Denys Lebed, Oleg K. Kolesnytskyj, Saule J. Rakhmetullina, Murat Mussabekov**
Automatic adjustment of reactive power by FACTS devices under conditions of voltage instability in the electric network
Automatyczna regulacja mocy biernej przez urządzenia FACTS w warunkach niestabilności napięcia w sieci elektrycznej 109

20. Petro Loboda, Ivan Starovit, Oleksii Shushura, Yevhen Havrylko, Maxim Saveliev, Natalia Sachaniuk-Kavets'ka, Oleksandr Neprytskyi, Dina Oralbekova, Dinara Mussayeva
Ventilation control of the new safe confinement of the Chornobyl nuclear power plant based on neuro-fuzzy networks
Kontrola wentylacji nowej bezpiecznej powłoki czarnobylskiej elektrowni jądrowej oparta na rozmytych sieciach neuronowych....114

21. Valeriy Kozlovskiy, Valeriy Kozlovskiy, Oleksii Nimych, Lyudmila Klobukova, Nataliia Yakymchuk
Model of the flat fairing antenna dielectric layer with aerodynamic heating
Model płaskiej warstwy dielektrycznej anteny z nagrzewaniem aerodynamicznym 119

22. Vitaly Pochernyaev, Nataliia Syvkova, Mariia Mahomedova
Microwave mixer on rectangular waveguides partially filled by dielectric
Mikser mikrofalowy na prostokątnych falowodach częściowo wypełnionych dielektrykiem 126

23. Liudmyla Matviichuk, Olena Liutak, Yuliia Dashchuk, Mykhailo Lepkiy, Svitlana Sidoruk
Information system for diagnostic competitiveness of the hospitality industry of the regions of Ukraine
System informatyczny dla diagnostyki konkurencyjności branży gospodarczej regionów Ukrainy....132

24. Oleksandr Harnaha, Natalia Savina, Volodymyr Hrytsiuk
Environmental and economic assessment of the land use regulation effectiveness
Środowiskowa i ekonomiczna ocena skuteczności rozporządzenia o użytkowaniu gruntów....139

List of reviewers cooperating with the journal in 2023
Lista recenzentów współpracujących z czasopismem w roku 2023 142

http://doi.org/10.35784/iapgos.5393 *received: 07.08.2023 | revised: 04.11.2023 | accepted: 29.11.2023 | available online: 20.12.2023*

A USAGE OF THE IMPEDANCE METHOD FOR DETECTING CIRCULATORY DISORDERS TO DETERMINE THE DEGREE OF LIMB ISCHEMIA

Valerii Kryvonosov[1], Oleg Avrunin[2], Serhii Sander[3], Volodymyr Pavlov[4], Liliia Martyniuk[1], Bagashar Zhumazhanov[5]

[1]National University of Life and Environmental Sciences of Ukraine, Kyiv, Ukraine, [2]Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, [3]Vinnitsia National Medical University named after M. I. Pirogov, Vinnytsia, Ukraine, [4]Vinnytsia National Technical University, Vinnytsia, Ukraine, [5]Institute of Information and Computing Technologies of the CS MES, Almaty, Kazahstan

Abstract. *New engineering technologies allow the creation of diagnostic devices for predicting the development of acute tissue ischemia of the extremities and determining the residual time until the removal of the tourniquet, and solving these tasks is particularly relevant during military actions. Acute limb ischemia is a sudden critical decrease in perfusion that threatens the viability of the limb. The incidence of this condition is 1.5 cases per 10 000 people per year. Acute ischemia occurs due to the blockage of blood flow in major arteries (embolism, thrombosis, trauma), leading to the cessation of adequate blood supply to metabolically active tissues of the limb, including the skin, muscles, and nerve endings. To address these issues, the article analyzes the changes in the impedance of biological tissue. The introduction and use of the coefficient of relative electrical conductivity, denoted as k, as a diagnostic criterion parameter, are justified. Experimental studies of changes in the coefficient of relative electrical conductivity k were conducted, confirming that the transition from exponential to linear dependencies of the coefficient establishes the degree of viability of the biological cell (tissue) and the moment of occurrence of reperfusion syndrome. It has been established that a deviation of the value of k by 10–15% from its unit value diagnoses the initial process of blood perfusion impairment and the development of ischemic tissue disease. The rate of change of k serves as a criterion for predicting the progression of the disease and as a corrective factor for therapeutic treatment.*

Keywords: ischemic tissue disease, perfusion, reperfusion syndrome, tourniquet, transient process, ionization

IMPEDANCYJNA METODA WYKRYWANIA ZABURZEŃ KRĄŻENIA KRWI DO OKREŚLENIA STOPNIA NIEDOKRWIENIA KOŃCZYNY

Streszczenie. *Nowe technologie inżynieryjne umożliwiają tworzenie urządzeń diagnostycznych do przewidywania rozwoju ostrego niedokrwienia tkanek kończyn i określania czasu pozostałego do usunięcia opaski uciskowej. Rozwiązanie tych zadań jest istotne, zwłaszcza podczas operacji wojskowych. Ostre niedokrwienie kończyny to nagłe, krytyczne zmniejszenie perfuzji, które zagraża żywotności kończyny. Częstość występowania tego schorzenia wynosi 1,5 przypadku na 10 000 osób rocznie. Do ostrego niedokrwienia dochodzi w wyniku zablokowania przepływu krwi w głównych tętnicach (zatorowość, zakrzepica, uraz), co prowadzi do ustania odpowiedniego ukrwienia metabolicznie aktywnych tkanek kończyny, w tym skóry, mięśni i zakończeń nerwowych. Aby rozwiązać wybrane problemy, w artykule przedstawiono analizę zmiany impedancji tkanki biologicznej. Uzasadnione jest wprowadzenie i stosowanie współczynnika przewodności elektrycznej względnej k, jako parametru kryterium diagnostycznego. Przeprowadzono eksperymentalne badania zmiany współczynnika względnego przewodnictwa elektrycznego k, które potwierdziły, że przejście zmiany współczynnika względnego przewodnictwa elektrycznego z zależności wykładniczej na liniową będzie determinować stopień żywotności komórki biologicznej (tkanki) i moment wystąpienia zespołu reperfuzyjnego. Ustalono, że odchylenie wartości k o 10–15% od jego wartości jednostkowej świadczy o początkowym procesie zaburzenia perfuzji krwi i rozwoju choroby niedokrwiennej tkanek, a szybkość zmiany jest kryterium do przewidywania rozwoju choroby i czynnika korygującego w postępowaniu terapeutycznym.*

Słowa kluczowe: choroba niedokrwienna tkanek, perfuzja, zespół reperfuzji, staza hemostatyczna, proces przejściowy, jonizacja

Introduction

The primary cause of death in trauma cases remains hemorrhage. The frequency of injuries to major vessels in modern military conflicts reaches 7.5–9.0% [3, 14]. Consequently, it is essential to address temporary hemostasis during the pre-hospital stage, leading to an increasing use of tourniquets [4, 5, 15]. Experimental studies [6, 17] have demonstrated that tourniquets can induce systemic changes. The application of a tourniquet (artificial limb ischemia) leads to the development of paraneuronal necrosis and reperfusion injuries not only in skeletal muscles but also in vital organs, such as the brain, heart, lungs, and kidneys. The most severe consequence of restoring arterial blood flow is reperfusion syndrome [27, 10].

According to [9, 19], the time a tourniquet can be applied depends on the patient's individual characteristics, environmental temperature, type of injury, and other factors, and can vary from 1 to 6 hours. The decision of when to remove the tourniquet is subject to the physician's judgment.

Disorders of blood circulation are also associated with a range of conditions, such as arterial and venous congestion (hyperemia), ischemia, blood stasis, thrombosis, disseminated intravascular coagulation (DIC) syndrome, embolism, infarction, bleeding, hemorrhage, diabetes mellitus, and others [11]. Currently, the frequency of blood circulation disorders is 1.5% per 10,000 people per year. Limb injuries resulting from occupational and domestic activities remain the leading cause of fatal outcomes and injuries among working-age individuals.

Blood circulation disorders lead to the development of necrosis and paraneuronal necrosis preceding necrobiosis. Acidosis occurs in tissues, and hyperkalemia develops [20].

Figure 1 illustrates the consequences of acute limb ischemia development.

Timely diagnosis of the initial process of tissue necrosis formation allows not only predicting the development of the disease but also adjusting the course of therapeutic treatment at early stages, thereby postponing surgical intervention and disability of the patient. In the case of hemostasis with a tourniquet, it helps determine the time remaining before the tourniquet is removed [3, 15, 25].

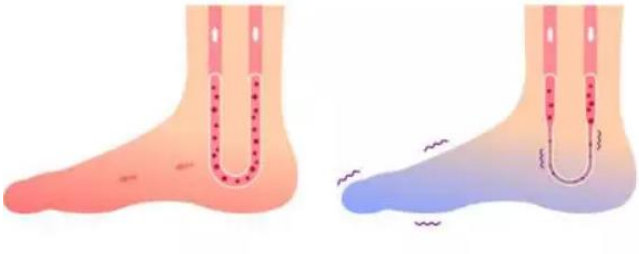


Fig. 1. Consequence of acute limb ischemia

1. Materials and methods of researsh

The main methods for diagnosing the development of ischemia in biological tissue during stationary research include differential diagnosis, ultrasound duplex scanning of lower limb vessels, computed tomographic angiography, magnetic resonance angiography, and digital subtraction angiography [14]. These methods require complex and expensive equipment, and they are typically used to refine the diagnosis and determine the extent of the disease. They are not cost-effective and may require a significant amount of time for diagnosis.

Electrical research methods on processes occurring in living organisms have been extensively studied [13, 17, 21] and widely applied. Many studies have focused on the effects of alternating, high-frequency currents on biological tissue.

For instance, in study [3, 9, 22], the influence of variable-frequency electric current on the electrical conductivity of skeletal muscles was investigated. The results presented in [19, 25] demonstrated changes in the electrical conductivity of biological structures under the influence of pulsed electric current in various frequency ranges. In [20, 28], methods of bioelectrical impedance were explored for non-invasive health monitoring using high-frequency alternating currents.

These research findings contribute to the understanding and application of electrical methods in the study and assessment of biological tissues and their responses to different electrical stimuli.

It should be noted that the utilization of changes in the electrical conductivity of biological tissues during the transitional period of ionization as an informative diagnostic criterion is not sufficiently studied in well-known scientific sources. As known, parane-crosis is manifested by alterations in the cytoplasm properties, shifting the reaction of cystazoa and the nucleus to the acidic side, release of potassium ions and phosphates from the cell, accumulation of sodium and chloride ions, increased colloid viscosity, nuclear compaction, enhanced sorption properties, and

overall changes in the electrolyte composition of biological tissues [8, 21, 30]. When applying a constant voltage to biological tissue, the main mechanism characterizing the flow of direct electric current at the initial moment is ionic conductivity. In study [6, 20, 28], it has been proven that the development of portable and cost-effective devices for studying the impedance changes in biological tissues has become significant at the present time.

The change in the property state of biological tissue during the development of necrosis correlates with changes in electrical conductivity. Therefore, by measuring the constant time of ionization during the transitional process of impedance change in biological tissue, it is possible to use it as an informative criterion for diagnosing ischemia. The duration of the transitional process of ionization in biological tissue, under the influence of direct current, does not exceed two minutes, which allows for a shorter diagnostic time and the development of cost-effective diagnostic devices.

Research objective: To justify the selection of a criterion diagnostic parameter describing changes in the properties of biological tissue (BT) under the influence of direct current voltage and formulate criteria for determining the critical state of BT properties. To develop a method for rapid diagnosis of changes in BT properties during ischemia development.

2. Model experiment

An electric method was applied – invasive measurement of the impedance of biological tissue (BT) using a constant voltage of 5 V. Sexually mature guinea pigs – males weighing 500–600 g were used for the study. The experiments were conducted under anesthesia, using thiopental sodium 10 mg/ml – 3-4 ml. Anesthesia was administered 30 minutes before the start of the experiment. During the 2-hour procedure, premedication was administered, consisting of diphenhydramine 1% – 0.3 ml and analgin 50% – 0.3 ml.

To assess the changes in BT impedance during circulatory disturbances, a subcutaneous ligature was applied to the right hind paw at the level of the hip joint. Tightening the ligature simulated the situation of applying a tourniquet. Measurements were taken simultaneously in areas with the tourniquet and without it [9, 10].

The research complied with the provisions of Article 26 of the Law of Ukraine No. 3447-VI dated October 16, 2012, "On the Protection of Animals from Cruelty" [1, 7, 15], and the requirements of the European Convention for the Protection of Vertebrate Animals Used for Experimental and Other Scientific Purposes [6, 15].

The needle electrodes used for measuring the impedance of biological tissue were placed in identical locations on each limb, with the needles oriented along the direction of the muscle tissue.

The experiments were conducted in the Photonics Laboratory of Vinnytsia National Technical University. Fig. 2 shows the laboratory setup for measuring the electrical conductivity of biological tissue.

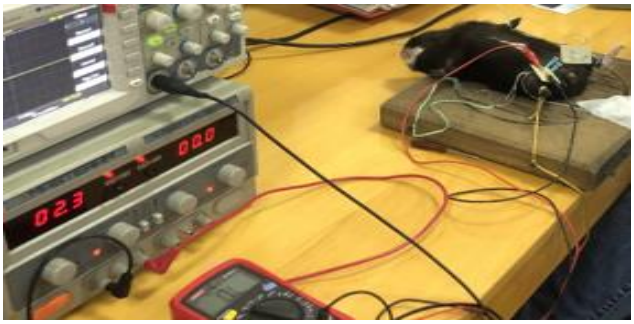


Fig. 2. Laboratory setup for measuring the impedance of BT

Medical needles with step-by-step distance adjustment were used as sensors for impedance measurements using the invasive method. The step adjustment was set at 1 cm. Fig. 3 presents the block diagram of the connection of the equipment for impedance measurements of biological tissue.

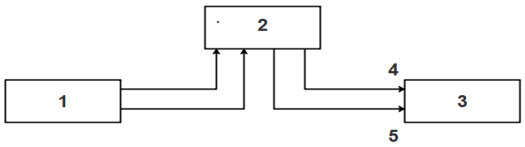


Fig. 3. Block diagram of the connection of equipment for impedance measurements of biological tissue, where indicated: 1 – DC voltage control unit; 2 – two-channel oscilloscope; 3 – biological object; 4 and 5 sensors for invasive conductivity measurement

For the analysis of the transient process, a mathematical model of biological tissue based on the corrected electrical equivalent circuit [12, 18, 24] was applied, which is presented in Fig. 4a and 4b. The equivalent circuit diagram (Fig. 4a), corresponds to the physico-chemical processes under the applied tourniquet, while the circuit in Fig. 4b corresponds to the investigation of healthy tissue.

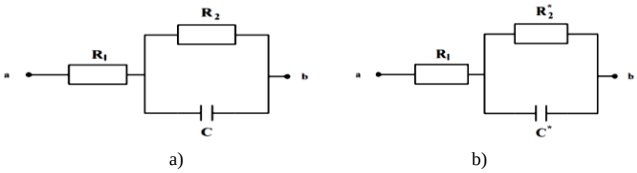


Fig. 4. Electrical equivalent circuits of biological tissue: a) without tourniquet, b) with applied tourniquet, where R_1 is the active resistance in the intercellular space, R_2 and R_2^* are the active resistances of tissue cells with and without the tourniquet, respectively, C and C are the capacitive components of cells characterizing the process of ionization in biological tissue, respectively*

Table 1. Results of the study of impedance changes over time depending on the distance between the needle electrodes

s [kΩ] / L [mm]	0	10	20	30	40	50	60	70	80	90	100	110	120
10	1.30	2.78	3.76	4.32	4.85	5.17	5.37	5.46	5.6	5.7	5.71	5.72	5.73
30	1.65	3.16	4.10	4.65	5.16	5.42	5.67	5.86	5.99	6.13	6.34	6.34	6.42
50	1.90	3.49	4.33	4.92	5.25	5.60	5.93	6.13	6.25	6.45	6.58	6.65	6.65

The mathematical model describing the change in resistance for the circuit (Fig. 4b) is as follows

$$Z = R_1 + R_2(1 - e^{-\frac{t}{\tau_1}}) \quad (1)$$

where $\tau_1 = C \cdot R_2$ the time constant of the ionization process for biological tissue without the applied tourniquet. Note that $R = \rho_0 \frac{L}{S}$, where ρ_0 – the specific resistance of tissue, L – the distance between the electrodes, S – the cross-sectional area, in this case, is constant.

To study the influence of the sensitivity of changes in the impedance of biological tissue on the distance between the electrodes for invasive measurements, the data obtained are presented in table 1.

Fig. 5 it shows the graphs of changes in the impedance of biological tissue during the transitional period, depending on the distance L between the needle electrodes, obtained experimentally. The investigated biological object was not harmed after the completion of the experiments.

The analysis of the curves (Fig. 5), indicates that the distance between the detector needles does not alter the nature of the transient process but only affects the sensitivity.

As known from studies [8, 9, 11], muscles contain potassium, sodium, calcium, magnesium, fluorine, and trace amounts of copper, manganese, zinc, cobalt, arsenic, and others. Among the anions, phosphoric and hydrochloric acids are found in the highest quantities. Muscles are characterized by a high content of potassium, phosphorus, and sulfur, and the primary component of the biological tissue (BT) environment is blood, which consists of electrolyte solutions. For instance, plasma contains 0.32% NaCa, with concentrations of Na+ ions at 142 mmol/L and K+ ions at 5 mmol/L. Electric current represents the directed movement of positive and negative ions and is determined according to [12] as:

$$j_+ = q_+ \cdot n_+ \cdot v_+ \text{ and } j_- = q_- \cdot n_- \cdot v_- \quad (2)$$

where q_+ – the positive charge carrier, q_- – the negative charge carrier; n_+, n_- – the amount of positively and negatively charged ions; v_+, v_- – the concentration of positively and negatively charged ions, respectively.

The total current will be equal to:

$$G = j_+ + j_- = q_+ \cdot n_+ \cdot v_+ + q_- \cdot n_- \cdot v_- \quad (3)$$

Expression (18) indicates that the electrical conductivity of biological tissue is proportional to n_+, n_- – the quantity of positively and negatively charged ions, and v_+, v_- – the concentration of positively and negatively charged ions, respectively.

The speed of ordered ion movement is directly proportional to E_{ion}, which is caused by the displacement of free charged ions under the influence of the electric field of the source and is determined as:

$$V = g \cdot E_{ion} \quad (4)$$

where g – the coefficient of proportionality of the carrier mobility.

The specific electrical conductivity δ for the electrolyte will be expressed as follows:

$$\delta = \frac{1}{z} = b_+ \cdot n_+ \cdot v_+ + b_- \cdot n_- \cdot v_- + b \cdot n \cdot v \quad (5)$$

The input impedance of the biological tissue for the circuit (Fig. 4b) with a limited number of ions, obtained by performing similar calculations as for the circuit (Fig. 4a), is determined by the following expression:

$$Z^*_{inp4} = R_1 + R^*_2(1 - e^{-\frac{t}{\tau^*}}) \quad (6)$$

where τ^*– the time constant of the ionization process for the biological object with a reduced number of ions is equal to: $\tau^* = C^* R^*_2$

The analysis of expressions (3), (4), and (5) shows that the time constant of the ionization process $\tau^* = C^* R^*_2$ is directly proportional to the number of charged ions, their concentration in the biological tissue, the level of electric field intensity E_{ion}, and the composition of salts in the biological tissue. It remains a constant value characterizing the flow of this process.

The specific electrical conductivity δ^* for the electrolyte is expressed as:

$$\delta^* = \frac{1}{Z^*} =$$

$$= b_+ \cdot (n_+ - \Delta n) \cdot v_+ + b_- \cdot (n_- - \Delta n) \cdot v_- + b \cdot (n - \Delta n) \cdot v \quad (7)$$

where $\Delta n = (n_+ - n^*_+)$ – the number of charged particles that did not enter the biological tissue due to perfusion.

Due to the increase in Δn, the following inequality is obviously satisfied:

$$Z^* < Z \; ; \quad \tau_2 < \tau_1 \quad (8)$$

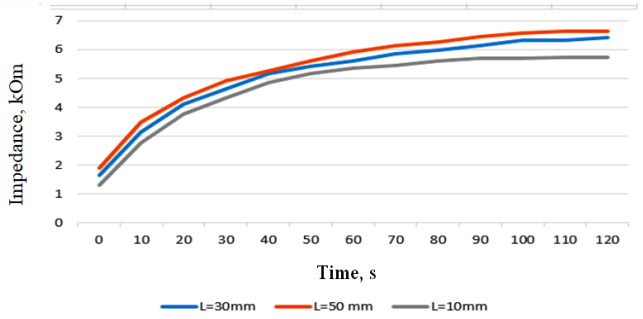


Fig. 5. Dependency of impedance change R = f(t) on the distance between the electrodes: L = 1 cm, L = 3 cm and L = 5 cm

3. Experimental results

Confirmation of expression (8) is provided by the presented results of measuring the change in tissue impedance at two extremities in identical locations. Fig. 6 shows the graph of tissue impedance variation over time during the cessation of blood circulation.

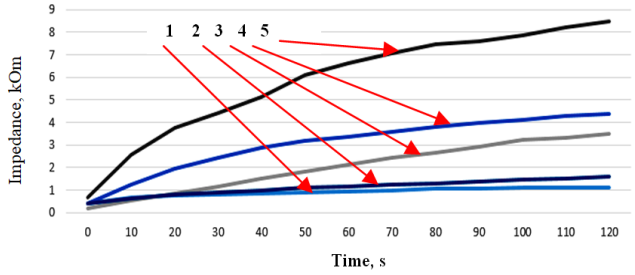


Fig. 6. Graph of tissue impedance change over time during the application of a tourniquet (1 – R,kOm (0.5 time), 2 – R,kOm (1.1 time), 3 – R,kOm (2.5 time), R,kOm (3.0 time), R,kOm (without tourniquet)

Fig. 7 shows the graph of changes in the ionization time constant with respect to the tourniquet application time.

The obtained relationship confirms that changes in the ionization time constant can be one of the diagnostic criteria for changes in the properties of biological tissue [16, 23, 26].

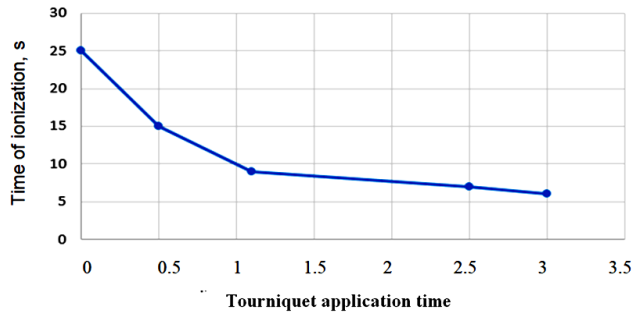


Fig. 7. Graph of changes in the ionization time constant as a function of tourniquet application time

The investigation of individual characteristics of impedance changes was conducted on four objects. Fig. 8 presents the graph of impedance variations for each specimen.

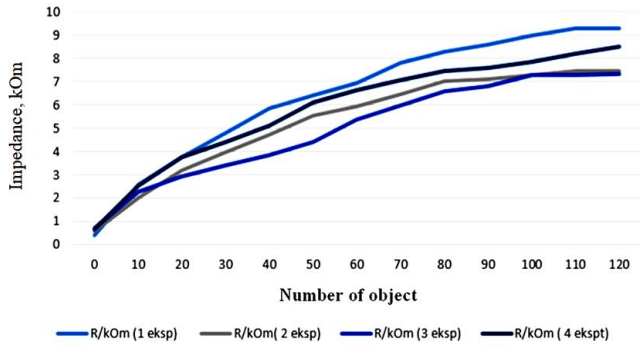


Fig. 8. Graph of impedance changes in biological tissue for four specimens

The analysis of the dependencies *R=f(t)* (Fig. 8), indicates that the variation in impedance values depends on the individual properties of the investigated specimens and ranges from 10% to 20%. In order to exclude individual human characteristics and determine diagnostic criteria for ischemia development that are independent of many uncertain factors, a diagnostic method has been developed as follows [2, 29, 30].

1. Electrodes are placed in pairs on corresponding areas of each limb.
2. Simultaneously measure the change in impedance of each limb segment.
3. The coefficient of relative impedance change is determined as follows: $k = \frac{Z^*}{Z}$, where Z^* – are the impedances of the biological tissue in the segments of the limb with and without the tourniquet applied, respectively. In cases of conditions such as diabetes or limb injury, the healthy limb is taken as the baseline value.

Fig. 9. The graphs of the coefficients of relative impedance change are presented as a function of the duration of tourniquet application.

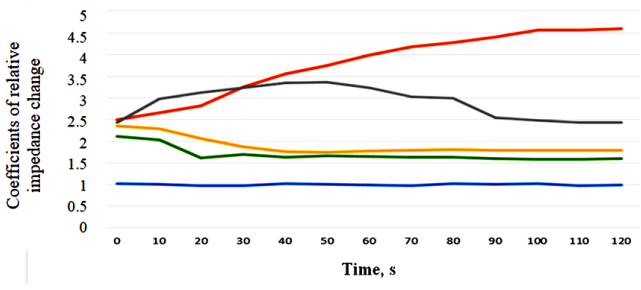


Fig. 9. Graphs of changes in the coefficients of relative impedance change as a function of the duration of tourniquet application

Analysis of the graphs, shown (Fig. 8), depicting changes in the relative impedance coefficient *k=f(t)* over time, allows for the establishment of criteria for assessing the condition and development of ischemia in biological tissue.

- the graph of the changes in k (0) for the investigation of two healthy limbs shows that the relative impedance coefficient remains within the range of 1 ± 0.05, fluctuating within 0.5% (the instrument's accuracy error). The value of k(0)=1±0,05 indicates an equal (healthy) state of the biological tissue.
- the graph of changes in k(1) during a 0.5-hour perfusion stop follows an exponential pattern, indicating the absence of biological changes in the tissue and its ability to recover.
- the coefficient k(1.1), corresponding to a 1.1-hour perfusion stop, changes according to a quadratic-decreasing law, characterizing the transition of the biological tissue from viable to the initial stages of tissue necrosis.
- the coefficients k(2.5) and k(3.0) obtained after a perfusion stop of 2.5 and 3.0 hours or more show minimal changes, indicating irreversible processes of tissue necrosis and the manifestation of the reperfusion effect.

On Fig. 10 graphs of changes in the impedance of BT after releasing the tourniquet are presented. The tourniquet was applied for 2.5 hours, and measurements were taken after its removal at 10 and 30 minutes.

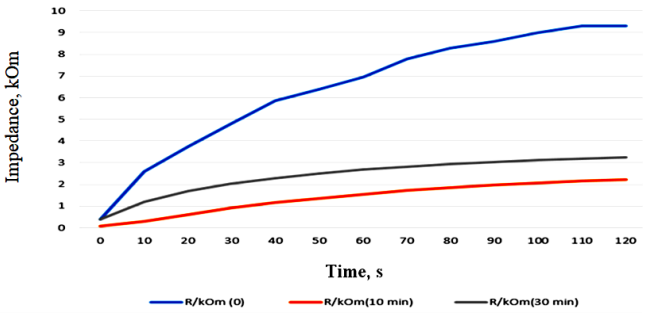


Fig. 10. Graphs of changes in the impedance of biological tissue after releasing the tourniquet

The analysis of the graphs (Fig. 10), provides grounds to consider that for the investigated biological object, the initial stage of cell necrosis and the onset of reperfusion syndrome are observed when blood perfusion is disrupted for more than 1.0 hour. The cessation of blood perfusion for 2.5 hours and more led to irreversible tissue necrosis processes. Tissue impedance practically does not recover and requires additional medical intervention.

4. Conclusions

Determining the remaining time until critical changes in the properties of biological tissue during the application of a tourniquet is not always possible in field conditions, considering the complexity of using existing stationary devices. Conducting rapid diagnosis of limb ischemia requires fast and cost-effective diagnostic tools, especially in the surgeon's office.

The study of the transient ionization process, which lasts up to 2 minutes, has shown that to avoid discrepancies related to individual patient characteristics, it is advisable to use the k-coefficients of relative impedance changes in biological tissue.

Based on the results of the study of the transient ionization process, a method for detecting changes in the properties of biological tissue during impaired blood circulation has been developed. The duration of the transient process is between 100 s and 120 s, allowing for rapid diagnosis of ischemia in the muscle

tissue of the patient's extremities. The method is based on comparing two simultaneously obtained integrated functions of healthy and affected tissues and analyzing the relative coefficient – *k*, which is determined by the formula: $k = \left(1 - \frac{S_{ex}}{S_{ex}^*}\right) \cdot 100\%$, where S_{ex} – the result of integrating the function *Z=f(t)* without the application of a blood occlusion cuff is denoted as S_ex, and the result of integrating the function *Z=f(t)* with the application of a blood occlusion cuff is denoted as S_{ex}^*. According to the data presented in Fig. 8, if the relative coefficient – *k* falls within the range of 0±0.05%, it indicates unchanged properties of biological tissues in both extremities. When the relative coefficient varies between 0.05% and 1, it suggests changes in the properties of biological tissues in both extremities. If the value of k is 0.5 or lower, it indicates the presence of irreversible processes in the development of ischemia in the muscle tissue.

It should be noted that the changes in the coefficient *k* according to an exponential law during the transitional process confirm the viability of biological tissue, allowing the avoidance of reperfusion syndrome after the removal of the blood occlusion cuff. On the other hand, changes in the relative electrical resistance coefficient k according to a linear law during the transitional process are associated with irreversible cell necrosis. Deviations in the value of the relative electrical resistance coefficient *k* by 5–8% from the unit value diagnose the initial stages of blood perfusion disruption and the development of ischemic tissue disease. The development of reperfusion syndrome is observed in the investigated biological objects when the perfusion is disturbed for more than 1 hour, and disruptions of perfusion exceeding 2.5 hours lead to irreversible changes in the properties of the biological tissue.

It has also been established that simultaneous measurements of impedance changes in two identical segments of limbs reveal the time constant of ionization as a critical parameter in studying the development of ischemia.

The prospects of the study – Advances in rapid diagnostic methods based on data obtained from transient processes of ionization in biological tissue require further research in conjunction with results from tissue property changes obtained through biopsy. This will allow the development of prediction methods for the development of ischemia and help determine the timing of surgical interventions and adjust the course of medical treatment.

Acknowledgments

Prepared with the grant support of the National Research Fund of Ukraine within the framework of the project 2022.01/0135 "Development of a laser-photon treatment and diagnostic complex for the medical rehabilitation of patients with polytraumas of various degrees of severity".

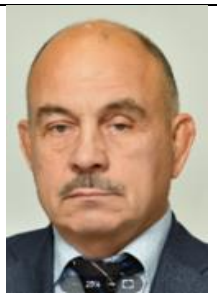
References

[1] Bera T. K.: Bioelectrical Impedance and The Frequency Dependent Current Conduction Through Biological Tissues: A Short Review. IOP Conference Series Materials Science and Engineering 331(1), 2018, 012005.
[2] Bera T. K.: Methods of bioelectrical impedance for non-invasive health monitoring. Review article. 2014, 381251.
[3] Blaisdell F. W.: The pathophysiology of skeletal muscle ischemia and the reperfusion syndrome: a review, Cardiovasc. Surg., 10(6), 2002, 620–630.
[4] Bosboom E. M., Hesselink M. K., Oomens C. W., Bouten C. V., Drost M. R., Baaijens F. P.: Passive transverse mechanical properties of skeletal muscle under in vivo compression. J Biomech, Oct 34(10), 2001, 1365–1368.
[5] Bouten C. V. C., Breuls R. G. M., Peeters E. A. G., Oomens C. W. J., Baaijens F. P. T.: In vitro models to study compressive strain-induced muscle cell damage. Biorheology 40(1–3), 2003, 383–388.
[6] Didukh V. D.: Biological physics with physical methods of analysis: teaching. Village, Ternopil, 2021.
[7] European Convention "On the Protection of Vertebrate Animals Used for Research and Other Scientific Purposes". Strasbourg, 1986.
[8] Frink M., Lechler P., Debus F., Ruchholtz S.: Multiple trauma and emergency room management. Dtsch. Arztebl. Int 114(29–30), 2017, 497–503.
[9] Jiang Z., Yao J., Wang L., Wu H., Huang J., Zhao T., Takei M.: Development of a portable electrochemical impedance spectroscopy system for bio-detection. IEEE Sensors Journal, 19(15), 2019, 5979–5987 [http://doi.org/10.1109/JSEN.2019.2911718].
[10] Katelyan O. V. et al.: Study of the peripheral blood circulation of an abdominal wall using optoelectronic plethysmograph. Information Technology in Medical Diagnostics II. CRC Press, Balkema book, Taylor & Francis Group, London, 2019, 119–125.
[11] Koutsouras D. A., Lingstedt L. V., Lieberth K., Reinholz J., Mailänder V., Blom P. W. M., Gkoupidenis P.: Probing the impedance of a biological tissue with PEDOT:PSS-Coated metal electrodes: Effect of electrode size on sensing efficiency. Adv. Healthcare Mater. 8, 2019, 1901215 [http://doi.org/10.1002/adhm.201901215].
[12] Kozlovska T. I. et al.: Device to determine the level of peripheral blood circulation and saturation. Proc. SPIE 10031, 2016, 100312Z.
[13] Krivonosov V. E., Pavlov S. V., Sander S. V., Martyniuk L. V.: Method of detection and control of the development of ischemia of biological tissue. Patent No. 118335, dated 11.05.2023.
[14] Kryvonosov V., Prudnikova N., Martyniuk L.: Justification of the electrical scheme of biological tissue replacementunder the action of DC voltage. Machinery & Energetics 13(4), 2022.
[15] Law of Ukraine No. 3447-VI "On the Protection of Animals from Cruelty", October 16, 2012.
[16] Maegele M., Spinella P., Schöchl H.: The acute coagulopathy of trauma: mechanisms and tools for risk stratification. Shock 38, 2012, 450–458.
[17] Mansoorifar A., Koklu A., Shihong M., Raj G. V., Beskok A.: Electrical Impedance Measurements of Biological Cells in Response to External Stimuli. Anal. Chem. 90(7), 2018, 4320–4327.
[18] Martsenyuk V. P.: Medical and biological physics. Ukrmedknyga, Ternopil, 2012.
[19] Naranjo-Hernández D., Reina-Tosina J., Min M.: Fundamentals, Recent Advances, and Future Challenges in Bioimpedance Devices for Healthcare Applications. 2019, 9210258 [http://doi.org/10.1155/2019/9210258].
[20] Norgren L., Hiatt W. R., Dormandy J. A., Nehler M. R., Harris K. A., Fowkes F. G.: Inter-Society Consensus for the Management of Peripheral Arterial Disease (TASC II). J Vasc Surg. 45, 2007, S5-S67.
[21] Oyeniyi B. T. et al.: Trends in 1029 trauma deaths at a level 1 trauma center. Injury 48(1), 2017, 5–12.
[22] Paradis S. et al.: Chronology of mitochondrial and cellular events during skeletal muscle ischemia-reperfusion. American Journal of Physiology. Cell Physiology, 310(11), 2016, C968–C982 [http://doi.org/10.1152/ajpcell.00356.2015].
[23] Pavlov S. V. et al.: Analysis of microcirculatory disorders in inflammatory processes in the maxillofacial region on based of optoelectronic methods. Przeglad Elektrotechniczny 93(5), 2017, 114–117.
[24] Pavlov S. V. et al.: Electro-optical system for the automated selection of dental implants according to their colour matching. Przeglad Elektrotechniczny 93(3), 2017, 121–124.
[25] Prasad A., Roy M.: Bioimpedance analysis of vascular tissue and fluid flow in human and plant body: A review. Biosystems Engineering 197, 2020, 170–187.
[26] Tereshchenko N. F., Tsapenko V. V., Chuhraev N. V.: Research of electrical conductivity of biological animals. Bulletin of NTUU "KPI". Instrumentation series 53(1), 2017.
[27] Wang G.: Holder David S: Electrical Impedance Tomography (1st edition). BioMedical Engineering OnLine 4, 2005, 27 [http://doi.org/10.1186/1475-925X-4-27].
[28] Wójcik W. et al.: Medical Fuzzy-Expert System for Assessment of the Degree of Anatomical Lesion of Coronary Arteries. International Journal of Environmental Research and Public Health 20(2), 2023, 979 [http://doi.org/10.3390/ijerph20020979].
[29] Wójcik W., Smolarz A.: Information Technology in Medical Diagnostics. CRC Press, 2017.
[30] Zarutskyi Y. L., Shudrak A. A.: Instructions for military field surgery. Chalchynska N. V.: Damage to the main vessels. Kyiv 2014, 351–373.

D.Sc. Valerii Kryvonosov
e-mail: yhtverf007@ukr.net

Doctor of Technical Sciences, professor, National University of Life and Environmental Sciences of Ukraine.
Field of scientific interests are connected with the method of diagnostics of control and management of electrical complexes.

http://orcid.org/0000-0002-8219-021X



D.Sc. Oleg Avrunin
e-mail: oleh.avrunin@nure.ua

Doctor of Technical Sciences, professor, Head of Biomedical Engineering Department, Kharkiv National University of Radio Electronics, Ukraine. Scientific supervisor of research work on research of theoretical and technical principles of diagnostics, assessment and correction of medical and social human conditions. Invited Professor in Gottfried Wilhelm Leibniz Universität Hannover (Germany) and Harbin Engineering University (China).

http://orcid.org/0000-0002-6312-687X

Ph.D. Sergii Sander
e-mail: sanderserg@gmail.com

Candidate of Medical Sciences, associate professor of the department of surgery of medical faculty No2 Vinnytsya National Medical University named after M. I. Pirogov. Field of scientific interests: diagnosis, prognosis and treatment of limb ischemia.

http://orcid.org/0000-0003-1667-6198

M.Sc. Volodymyr Pavlov
e-mail: machinehead6926@gmail.com

Scientific researcher Vinnytsia National Technical University.
Scientific direction: biomedical information optoelectronic and laser technologies for diagnostics and physiotherapy influence.
Deals with issues of improving the distribution of optical radiation theory in biological objects, particularly through the use of electro-optical systems, and the development of intelligent biomedical optoelectronic diagnostic systems.

http://orcid.org/0000-0002-0717-7082

M.Sc. Liliia Martyniuk
e-mail: martyniuklilia@nubip.edu.ua

Senior lecturer of the Department of Electrical Supply named after V. M. Synkova of Education and research institute of Energetics, Automatics and Energy saving National University of Life and Environmental Sciences of Ukraine.

http://orcid.org/0009-0007-3852-5610

Ph.D. Bagashar Zhumazhanov
e-mail: bagasharj@mail.ru

Senior researcher of the Institute of Information and Computing Technologies of the CS MES RK, associate professor.
Graduated from the Faculty of Physics and Mathematics of the Abai Almaty State University with a degree in Computer Science and Mathematics. Master's degree in "Information Systems". In 2011 he defended his thesis for the degree of Candidate of Technical Sciences. Associate professor of the OP "Digital Engineering and Data Analysis".

http://orcid.org/0000-0002-5035-9076

http://doi.org/10.35784/iapgos.5380 received: 31.07.2023 | revised: 01.10.2023 | accepted: 14.11.2023 | available online: 20.12.2023

USAGE OF ARTIFICIAL NEURAL NETWORKS IN THE DIAGNOSIS OF KNEE JOINT DISORDERS

Konrad Witkowski[1], Mikołaj Wieczorek[2]
[1]Lodz University of Technology, Lodz, Poland, [2]Synerise S.A., Krakow, Poland

Abstract. *Following article address the issue of automatic knee disorder diagnose with usage of neural networks. We proposed several hybrid neural net architectures which aim to successfully classify abnormality using MRI (magnetic resonance imaging) images acquired from publicly available dataset. To construct such combinations of models we used pretrained Alexnet, Resnet18 and Resnet34 downloaded from Torchvision. Experiments showed that for certain abnormalities our models can achieve up to 90% accuracy.*

Keywords: classification, MRI images, Resnet, Alexnet

ZASTOSOWANIE SZTUCZNYCH SIECI NEURONOWYCH W DIAGNOZIE SCHORZEŃ STAWU KOLANOWEGO

Streszczenie. *Niniejszy artykuł porusza temat automatycznej diagnozy uszkodzenia stawu kolanowego z zastosowaniem sieci neuronowych. Zaproponowano kilka hybrydowych sieci neuronowych, które podjęły próbę poprawnej klasyfikacji nieprawidłowości wykorzystując zdjęcia rezonansu magnetycznego pochodzące z publicznie dostępnego zbioru. Do konstrukcji kombinacji sieci skorzystano z pretrenowanych modeli (Alexnet, Resnet18, Resnet34) pobranych z Torchvision. Eksperyment pokazał, że dla klasyfikacji niektórych schorzeń modele osiągnęły nawet 90% skuteczności.*

Słowa kluczowe: klasyfikacja, zdjęcia MRI, Resnet, Alexnet

Introduction

Knee joint disorders are a problem strictly combined with the human aging process. Such disorders are outcomes of everyday work and accidents that lead to physical damage. One of the most effective diagnose methods of such injuries is the analysis of MRI images.

In this project we tried to construct a hybrid neural network architecture that could possibly accurately classify knee joint abnormalities using MRI images uploaded by Stanford University as "A Knee MRI Dataset And Competition" [8]. The researchers from Stanford ML Group also published an article [1] presenting results of their models which served as a reference point to scores achieved by our neural nets. We would also acknowledge the fact that for better understanding of our task we analysed Ahmed Besbes's implementation available here [5].

The goal of the whole project is to check whether a single person's exam consisting of 3 planes (axial, coronal and sagittal) indicate the occurrence of an injury like abnormality, ACL tears or meniscal tears. Each exam was viewed and tagged with labels by medical doctors.

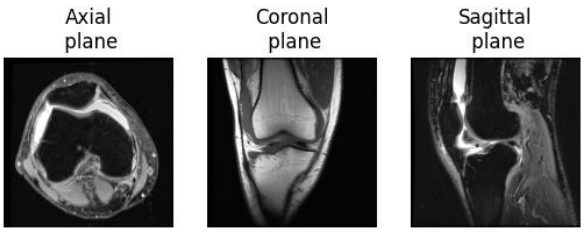


Fig. 1. Single exam's planes

This article is constructed as follows. At the beginning of the text we describe the dataset and the idea standing behind the experiment. After that we present the types of used neural nets and statistical methods. The last part of the paper is dedicated for experiment's results and summary.

1. Dataset characteristics

The dataset consists of 1 370 MRI examinations taken at Stanford University Medical Center. Each of the examination has 3 labels indicating presence of abnormality, ACL tears and meniscal tears. Occurrence of ACL tears or meniscal tears means that the abnormality label will be positive but it doesn't work the other way round. That means that the abnormality label covers not only ACL tears and meniscal tears but also other types of abnormalities not specified among labels.

MRI images were taken using various devices (GE Discovery, GE Healthcare, Waukesha, WI). Moreover two types of magnetic fields were used: 3.0 T (55.6% of exams) and 1.5 T for the rest of the exams.

Data uploaded by Stanford University was already pre-processed. That included converting DICOM (Digital Imaging and Communications in Medicine) files to png format and rescaling them to 256×256 resolution. Given that the images didn't have the same pixel intensity the researchers used standardization algorithm which based on pixel intensity taken from training dataset. The algorithm itself was run on both training and testing dataset.

In order to enhance the training dataset we performed augmentation consisting of random rotation, transposition and horizontal flip.

Table 1. Labels frequency

Label	Number of occurences
Abnormality	1 104
ACL Tears	319
Meniscal Tears	508

2. Experiment description

For each plane we attempted to build a single model (called submodel) which specialized in a specific label. To boost the performance of models we decided to use pretrained versions of Resnet18, Resnet34 and Alexnet downloaded from Torchvision [7] which served us as main parts of our submodels. Overview of their structures are available here: [2, 3]. The idea standing behind single net's functioning was to process the outcome of the pretrained model with average and max pooling, concatenate the results and finally perform calculations using fully connected layer. Whole structure is presented in Fig. 2.

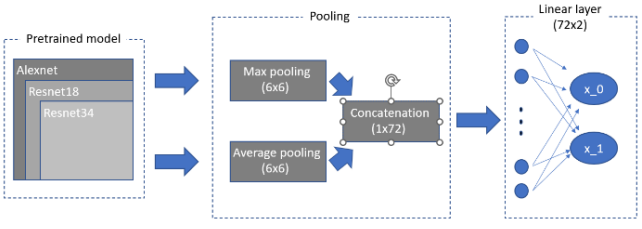


Fig. 2. Submodel's architecture scheme

The best suited model for classifying presence of specific label using given plane was chosen based on its performance and the results of McNemar's test run between all models.

The main models were composed out of 3 submodels each taking as input specific plane. Outcome of each submodel was sent to linear layer which ended up with 2 neurons.

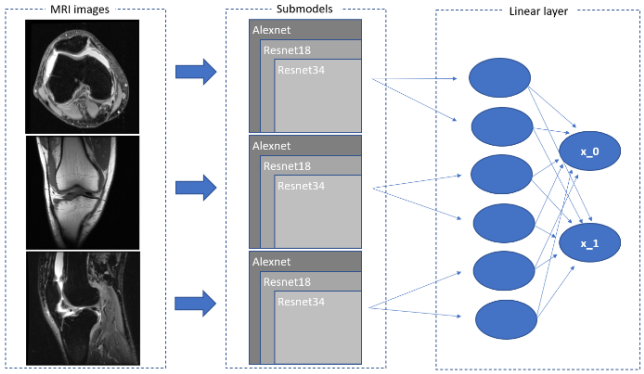


Fig. 3. Main models' architecture scheme

All neural nets were trained using binary cross entropy loss given by:

$$l = -w[y * ln\sigma(x) + (1-y) * \ln(1-\sigma(x)) \quad (1)$$

where x stands for predicted value, y for actual value, σ for logit function and w for weight loss.

To balance unequal label distribution we multiplied losses from actual positive observations with reversed proportion of number of actual positive observations to number of actual negative observations. This operation can be describes as follows:

$$w = \begin{cases} \frac{n.of\ actual\ negative\ obs.}{n.of\ actual\ positive\ obs.}, & y = 1 \\ 1, & y = 0 \end{cases} \quad (2)$$

3. Resnet

Resnet is a type of a neural network created by a group of researchers from Microsoft. Their main objective was solving the issue of degradation which takes place at the training process. The symptoms of this phenomenon appeared as deteriorated loss values not only on training set but also on test set whenever the construction of neural net was expanded with extra layers.

As a result, the researchers invented residual block which at that time differed from standard neural networks with the idea of using input vector at the beginning and at the end of set of layers.

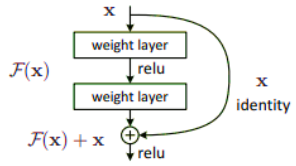


Fig. 4. Residual block scheme

Microsoft's researchers formulated a hypothesis saying that it's possible in an asymptotic way to estimate the outcomes of complicated functions. In the case of Resnet the authors took a step further and checked whether couple of connected layers are able to estimate outcome of residual function:

$$F(\boldsymbol{x}) = \mathrm{H}(\mathbf{x}) - \boldsymbol{x} \quad (3)$$

where H(x) is a covered mapping. It's possible to describe the way of working of residual block depicted in Fig. 4 in a following way:

$$\boldsymbol{y} = F(\mathbf{x}, \{W_i\}) + \boldsymbol{x} \quad (4)$$

where **x**, **y** are input and output tensor from the residual block. W_i describes the weights of *i*-th layer. The full version of *F* function can be expanded with 2 layers visible in Fig. 4 which gives:

$$F = W_2 \cdot \sigma(W_1 \boldsymbol{x}) \quad (5)$$

where σ is Relu activation function. The transformation presented in the equation (5) is called the skip connection and works only when dimensions of x and F are equal. When it's not the case then linear projection W_s is needed:

$$\boldsymbol{y} = F(\mathbf{x}, \{W_i\}) + \boldsymbol{x}W_s \quad (6)$$

Resnet's architecture is in fact a stack of residual blocks with implemented skip connections.

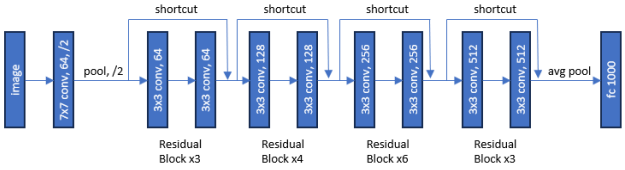


Fig. 5. Resnet34's architecture scheme

4. Alexnet

Alexnet is a type of neural network created by Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. Its success was based on several of innovations, whose Alexnet's creators were not always authors of, that were used all together in one architecture. Alexnet's training was spread around 2 graphic cards which allowed updating of 2 parallel mapping series and efficient memory management. Alexnet's structure begins with 3 convolutional layers which share data between graphic cards – input mapping for each of layer is structured from output tensors created by the previous layers placed on both graphic cards. Next 2 layers are convolutional type but in this case they are independent in sense of data sharing. They are followed by 3 fully connected layers.

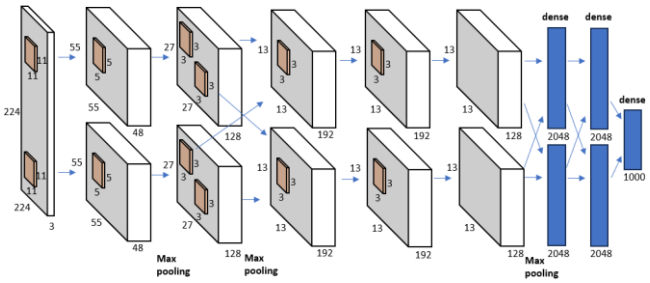


Fig. 6. Alexnet's architecture scheme

One of the most groundbreaking innovation was activation function called Relu which was described by following equation:

$$f(x) = \max(0, x) \quad (7)$$

At that time most of the neural networks used hyperbolic tangent which led to slower training tempo.

Another idea implemented in Alexnet was the local response normalization inspired by a phenomenon called the lateral inhibition which is characterized by excited neuron disabling its neighbors. This phenomenon leads to specialization of overexcited neurons in detecting certain patterns. Formally the local response normalization looks in the following way:

$$b^i_{x,y} = \frac{a^i_{x,y}}{(k + \alpha \sum_{j=\max(0,\ i-m/2)}^{\min(N-1,\ i+\frac{n}{2})} (a^j_{x,y})^2)^\beta} \quad (8)$$

where: $a^i_{x,y}$ is the outcome of applying *i*-th filter on element placed at (*x*,*y*) position and Relu function, $b^i_{x,y}$ is the normalized response for element placed at position (x,y) after applying *i*-th filter and *k*, *n*, *α*, *β* are the hyperparameters.

5. McNemar's test and Wilson's confidence interval

For comparison of models' classification results we performed McNemar's test whose overview is available here [6] and here [4]. This method uses convergence table which splits the same observations classified by two compared models into 4 groups presented in Table 2.

Table 2. Convergence table

	Classifier 1 correct results	Classifier 1 incorrect results
Classifier 1 correct results	a	b
Classifier 1 incorrect results	c	d

The null hypothesis states that both classifiers disagree to the same extend. If the null hypothesis is rejected, it means that there's a possibility that both classifiers disagree in a different way. To perform McNemar's test p value should be calculated which, depending on alfa value (here 0.05), confirms or rejects the null hypothesis:

p value > 0.05 → null hypothesis confirmed,

p value ≤ 0.05 → null hypothesis rejected.

In many cases chi square distribution is impossible to estimate since b + c < 25. Because of that reason we decided to use the exact p value given by the following equation:

$$exact\ value\ p = 2 \cdot \sum_{i=b}^{n} \binom{n}{i} \cdot 0.5^i \cdot (1-0.5)^{n-i} \quad (9)$$

where n = b + c.

In order to estimate the values of expected metrics on unseen before data we calculated Wilson confidence interval. A comprehensive overview is available here [9]. This method allowed us to construct range of expected results with given probability (95% in this project). Wilson confidence interval is given by the following equation:

$$p \approx (w^-, w^+) = \frac{1}{1+\frac{z^2}{n}} \cdot (\hat{p} + \frac{z^2}{2n}) \pm \frac{z}{1+\frac{z^2}{n}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}} \quad (10)$$

where n is the number of observations, z is z score for 95% confidence interval and $\hat{p}$ is the number of positive observations.

6. Results of submodels

For each label, for each plane, and for each pretrained model we performed a training process which lasted 10 epochs. Among 10 checkpoints we selected the one which obtained the highest accuracy on test set at the end of epoch. If at least 2 checkpoints achieved the same accuracy, then we chose the one which had the highest AUC (area under curve) result. At the end of the selection process we ended up with 27 submodels which we had to reduce to 9 – one submodel for (plane, label) pair.

To analyze submodel's performance we calculated the following statistics: accuracy, precision, recall, F1 score and AUC. To take a deeper look into delivered statistics we computed Wilson confidence interval. Furthermore we calculated p values using McNemar's test to find out whether submodels are statistically different.

Tables 3 and 4 present an example of results obtained by 3 submodels dedicated for abnormality classification using axial plane. In this case because of the high p values we decided to move forward with Resnet18 which has the least number of parameters.

Table 3. Classification of abnormality using axial plane

Model	Accuracy	Precission	Recall	F1	AUC
Alexnet	0.83 (0.753, 0.887)	0.86 (0.787, 0.911)	0.94 (0.882, 0.97)	0.9 (0.833, 0.942)	0.688 (0.601, 0.764)
Resnet18	**0.87 (0.798, 0.919)**	**0.87 (0.798, 0.919)**	**0.98 (0.936, 0.994)**	**0.92 (0.857, 0.956)**	**0.709 (0.623, 0.783)**
Resnet34	0.86 (0.787, 0.911)	0.86 (0.787, 0.911)	0.98 (0.936, 0.994)	0.92 (0.857, 0.956)	0.689 (0.602, 0.756)

Table 4. p values between submodels classifying abnormality using axial plane

Models		p value
Resnet18	Resnet34	1
Resnet18	Alexnet	0.424
Resnet34	Alexnet	0.648

We would also like to present insight into training process of chosen model. Fig. 7 and Fig. 8 show us that, even though all models were pretrained beforehand, the loss levels reached during training and testing looked different for submodel equipped with Alexnet and its equivalents with Resnets. The Alexnet submodel needed much more time to reach loss level represented by Resnet submodels.

Using the same strategy as described in the given example we selected the rest of submodels whose overview is presented in Tables 5 and 6.

Table 5. Final submodels chosen to build main models

Model	Accuracy	Precission	Recall	F1	AUC
Abnormality					
Resnet18 (axial)	0.87 (0.798, 0.919)	0.87 (0.798, 0.919)	0.98 (0.936, 0.994)	0.92 (0.857, 0.956)	0.709 (0.623, 0.783)
Resnet18 (coronal)	0.82 (0.742, 0.878)	0.85 (0.775, 0.903)	0.95 (0.895, 0.977)	0.9 (0.833, 0.942)	0.654 (0.565, 0.733)
Resnet18 (sagittal)	0.82 (0.742, 0.878)	0.82 (0.742, 0.878)	0.98 (0.936, 0.994)	0.89 (0.821, 0.934)	0.589 (0.5, 0.673)
ACL tears					
Resnet18 (axial)	0.75 (0.666, 0.819)	0.85 (0.775, 0.903)	0.54 (0.451, 0.627)	0.66 (0.571, 0.739)	0.731 (0.645, 0.802)
Resnet34 (coronal)	0.85 (0.775, 0.903)	0.86 (0.787, 0.911)	0.8 (0.72, 0.862)	0.83 (0.753, 0.887)	0.845 (0.77, 0.899)
Resnet34 (sagittal)	0.83 (0.753, 0.887)	0.83 (0.753, 0.887)	0.8 (0.72, 0.862)	0.81 (0.731, 0.87)	0.83 (0.753, 0.887)

Table 6. Final submodels chosen to build main models

Model	Accuracy	Precission	Recall	F1	AUC
Meniscal tears					
Resnet18 (axial)	0.67 (0.582, 0.748)	0.62 (0.531, 0.702)	0.62 (0.531, 0.702)	0.62 (0.531, 0.702)	0.661 (0.572, 0.739)
Resnet18 (coronal)	0.78 (0.698, 0.845)	0.75 (0.666, 0.819)	0.73 (0.644, 0.801)	0.74 (0.655, 0.81)	0.77 (0.687, 0.836)
Resnet34 (sagittal)	0.72 (0.634, 0.793)	0.72 (0.634, 0.793)	0.6 (0.511, 0.683)	0.65 (0.561, 0.729)	0.71 (0.623, 0.784)

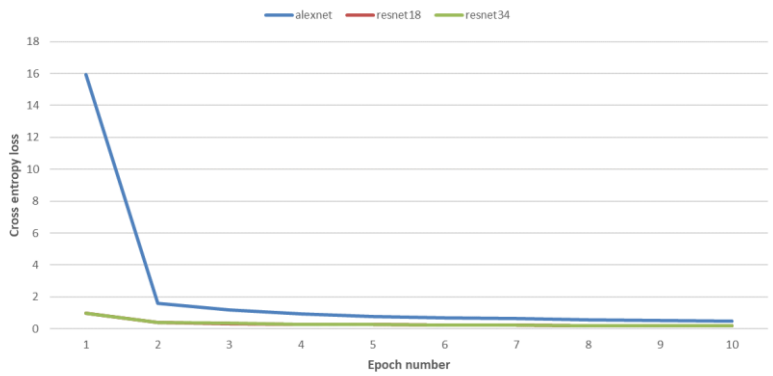


Fig. 7. Loss levels of submodels (classifying abnormality on train dataset) at the end of each epoch

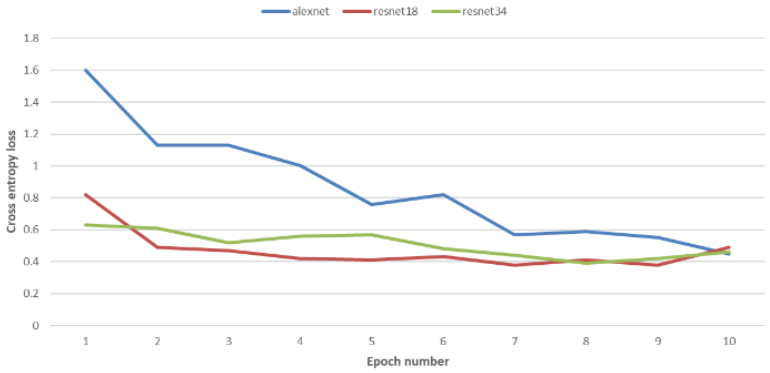


Fig. 8. Loss levels of submodels (classifying abnormality on test dataset) at the end of each epoch

7. Results of main models

To assess main models' effectiveness we decided to calculate the same metrics as in the submodels' cases but this time we expanded the analysis with specificity.

Table 7. Final models comparison compared with equivalents from Stanford University

Model	Accuracy	Precission	Specificity	Recall	F1	AUC
Abnormality						
Stanford's model	0.85 (0.775, 0.903)	No data	0.88 (0.800, 0.929)	0.71 (0.500, 0.862)	No data	0.94 (0.895, 0.937)
Authorial model	0.87 (0.798, 0.919)	0.86 (0.787, 0.911)	0.99 (0.952, 0.998)	0.4 (0.317, 0.489)	0.92 (0.857, 0.956)	0.69 (0.607, 0.77)
ACL tears						
Stanford's model	0.87 (0.794, 0.916)	No data	0.759 (0.635, 0.850)	0.97 (0.890, 0.991)	No data	0.97 (0.938, 0.965)
Authorial model	0.9 (0.833, 0.942)	0.9 (0.833, 0.942)	0.87 (0.798, 0.919)	0.924 (0.862, 0.959)	0.88 (0.81, 0.927)	0.9 (0.83, 0.94)
Meniscal tears						
Stanford's model	0.725 (0.639, 0.797)	No data	0.710 (0.587, 0.808)	0.74 (0.616, 0.837)	No data	0.85 (0.78, 0.847)
Authorial model	0.77 (0.687, 0.836)	0.69 (0.602, 0.766)	0.85 (0.775, 0.903)	0.706 (0.619, 0.78)	0.76 (0.676, 0.828)	0.78 (0.694, 0.841)

In each of 3 pairs of compared models the same pattern can be spotted – the differences in accuracy between Stanford's models and authorial models are low. For instance accuracy of Stanford's model responsible for classifying images with abnormality presence reached 0.85 in comparison to authorial model's 0.87.

It seems that the neural nets created by Stanford ML Group work much better in detecting those images which don't have sought label. In contrast to them authorial models excel in finding disorders in the images that actually present joint with disorder. For example recall and specificity of Stanford's model classifying ACL tears reached levels of 0.759 and 0.924. The same metrics for authorial model stood at 0.9 and 0.924.

It's worth mentioning that Stanford's neural networks overtook the authorial models when it comes to the AUC levels. The possible explanation for that could be the difference in a way of selecting submodel's checkpoint among epochs during training. The original Stanford's paper mentions that the researched chose those versions which had the lowest averaged loss counted within epoch. On the other hand authorial submodels were picked according to the highest accuracy.

8. Summary

In conclusion we would like to say that the created models that served to classify 3 types of knee joint disorder achieved comparable results as their equivalents from Stanford University. Their differences in a way of selection are with no doubts a good material for further research.

It seems that the topic of classifying knee joint injuries using neural nets is worth spending much more time on it. In our opinion aspects like choice of pretrained model or the construction of submodel could be much better explored. It's also clear to us that such models should help the medical doctors not only in the proper classification but also In pointing the place where injury is located. This was the main idea of Researchers from Stanford University who implemented class activation mapping – a heatmap generating technique which shows which part of the image were significant in classification.

References

[1] Bien N. et al.: Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Med 15(11), 2018, e1002699 [http://doi.org/10.1371/journal.pmed.1002699].
[2] He K., Zhang X., Ren S., Sun J.: Deep Residual Learning for Image Recognition. Computer Vision and Pattern Recognition 2015, arXiv:1512.03385.
[3] Krizhevsky A., Sutskever I., Hinton G. E.: ImageNet Classification with Deep Convolutional Neural Networks. F. Pereira, C. J. Burges, L. Bottou and K. Q. Weinberger: Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012.
[4] https://en.wikipedia.org/wiki/McNemar%27s_test
[5] https://github.com/ahmedbesbes/mrnet
[6] https://machinelearningmastery.com/mcnemars-test-for-machine-learning/
[7] https://pytorch.org/vision/stable/models.html
[8] https://stanfordmlgroup.github.io/competitions/mrnet/
[9] https://www.mikulskibartosz.name/wilson-score-in-python-example/

M.Sc. Konrad Witkowski
e-mail: k.l.p.witkowski@gmail.com

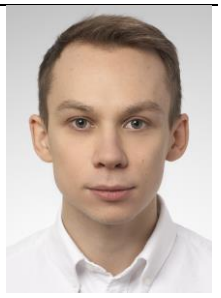
Konrad Witkowski received his M.Sc. from the SGH Warsaw School of Economics. His research interest include neural nets, machine learning etc.

http://orcid.org/0009-0004-2916-8672

M.Sc. Mikołaj Wieczorek
e-mail: mwieczorek.doc@gmail.com

Mikołaj Wieczorek, holds a Master's in Operational Research with Data Science from the University of Edinburgh. Professionally works as ML/MLOps Engineer. His expertise lies in cloud infrastructure, AI model deployment and optimisation. His research interest include Computer Vision, NLP and recommender systems.

http://orcid.org/0000-0003-4397-3331



http://doi.org/10.35784/iapgos.5309 *received: 16.09.2023 | revised: 10.11.2023 | accepted: 26.11.2023 | available online: 20.12.2023*

COMPREHENSIVE MACHINE LEARNING AND DEEP LEARNING APPROACHES FOR PARKINSON'S DISEASE CLASSIFICATION AND SEVERITY ASSESSMENT

Oumaima Majdoubi, Achraf Benba, Ahmed Hammouch
Mohammed V University in Rabat, National School of Arts and Crafts, Electronic Systems Sensors and Nanobiotechnology, Rabat, Morocco

Abstract. *In this study, we aimed to adopt a comprehensive approach to categorize and assess the severity of Parkinson's disease by leveraging techniques from both machine learning and deep learning. We thoroughly evaluated the effectiveness of various models, including XGBoost, Random Forest, Multi-Layer Perceptron (MLP), and Recurrent Neural Network (RNN), utilizing classification metrics. We generated detailed reports to facilitate a comprehensive comparative analysis of these models. Notably, XGBoost demonstrated the highest precision at 97.4%. Additionally, we took a step further by developing a Gated Recurrent Unit (GRU) model with the purpose of combining predictions from alternative models. We assessed its ability to predict the severity of the ailment. To quantify the precision levels of the models in disease classification, we calculated severity percentages. Furthermore, we created a Receiver Operating Characteristic (ROC) curve for the GRU model, simplifying the evaluation of its capability to distinguish among various severity levels. This comprehensive approach contributes to a more accurate and detailed understanding of Parkinson's disease severity assessment.*

Keywords: Parkinson's disease, severity assessment, machine learning, XGBoost, Gated Recurrent Unit (GRU), comparative analysis

KOMPLEKSOWE METODY UCZENIA MASZYNOWEGO I UCZENIA GŁĘBOKIEGO DO KLASYFIKACJI CHOROBY PARKINSONA I OCENY JEJ NASILENIA

Streszczenie. *W tym badaniu naszym celem było przyjęcie kompleksowego podejścia do kategoryzacji i oceny ciężkości choroby Parkinsona poprzez wykorzystanie technik zarówno uczenia maszynowego, jak i głębokiego uczenia. Dokładnie oceniliśmy skuteczność różnych modeli, w tym XGBoost, Random Forest, Multi-Layer Perceptron (MLP) i Recurrent Neural Network (RNN), wykorzystując wskaźniki klasyfikacji. Wygenerowaliśmy szczegółowe raporty, aby ułatwić kompleksową analizę porównawczą tych modeli. Warto zauważyć, że XGBoost wykazał najwyższą precyzję na poziomie 97,4%. Ponadto poszliśmy o krok dalej, opracowując model Gated Recurrent Unit (GRU) w celu połączenia przewidywań z alternatywnych modeli. Oceniliśmy jego zdolność do przewidywania nasilenia dolegliwości. Aby określić ilościowo poziomy dokładności modeli w klasyfikacji chorób, obliczyliśmy wartości procentowe nasilenia. Ponadto stworzyliśmy krzywą charakterystyki operacyjnej odbiornika (ROC) dla modelu GRU, upraszczając ocenę jego zdolności do rozróżniania różnych poziomów nasilenia. To kompleksowe podejście przyczynia się do dokładniejszego i bardziej szczegółowego zrozumienia oceny ciężkości choroby Parkinsona.*

Słowa kluczowe: choroba Parkinsona, ocena ciężkości, uczenie maszynowe, XGBoost, Gated Recurrent Unit (GRU), analiza porównawcza

Introduction

Parkinson's disease (PD) is a progressive neurological condition that deteriorates as it progresses, primarily because of the untimely demise of dopaminergic neurons in the substantia nigra area [17]. Individuals are affected in their fundamental physiological systems, including respiration, balance, movement, and heart function [22]. In the early stages, their ability to speak fluently is hindered. Early detection of PD leads to extended patient lifespans, and accurate diagnosis requires robust health informatics tools. These solutions are created to assist healthcare professionals [7, 13] in assessing PD severity by utilizing diverse sensors. Symptoms of PD are linked to disturbances in dopaminergic pathways, where there is a deficiency of dopaminergic neurons, it results in a combination of motor and non-motor symptoms. Motor manifestations consist of trembling, rigidity, reduced mobility, and gait difficulties. On the other hand, non-motor symptoms encompass mood disorders, hallucinations, and incidents, urinary and reproductive issues, as well as sleep disturbances [5]. These symptoms emerge when approximately 60% of dopaminergic neurons remain, and they correlate with the effects of aging, ultimately diminishing quality of life [3]. This research supports healthcare professionals in assessing the severity of Parkinson's disease by employing various sensors. The approach involves utilizing diverse speech signal processing methods to capture vital attributes associated with Parkinson's disease. Subsequently, these attributes undergo analysis through machine learning algorithms, providing both dependable PD detection and an evaluation of disease severity. This information can facilitate early intervention and treatment strategies.

Recent years have witnessed a surge in the utilization of deep learning methods in medical research, showcasing their capacity to unravel intricate patterns from intricate data. This advancement enhances disease diagnosis accuracy and dependability. In this context, our research investigates the effectiveness of deep learning models. This entails utilizing contemporary neural network structures like MLP and RNN, in combination with established machine learning methods such as XGBoost and Random Forest, to improve the classification and severity evaluation of Parkinson's disease (PD).

Our research aims to bridge the gap between clinical expertise and computational prowess, presenting an encompassing framework. This framework not only precisely categorizes PD but also quantifies its severity. We tap into a well-curated dataset containing an array of patient features, executing a multifaceted strategy encompassing data pre-processing, exploratory analysis, model selection, regularization methods, and dimensionality reduction. By evaluating diverse models based on crucial metrics like precision, recall, and F1-score, we seek to comprehend the strengths and limitations of each approach concerning PD classification. Additionally, our study incorporates a GRU model to compute severity assessment percentages, enhancing the comprehensiveness of our analysis.

In this section, we will explore ongoing efforts in the classification of Parkinson's disease (PD). Using machine learning techniques and explore recent advancements in deep learning methods for the same purpose. Additionally, we will investigate the utilization of speech recordings not only for PD classification but also for evaluating disease severity. This entails using advanced computational models to categorize individuals based on PD presence and quantify the severity, offering crucial insights for personalized medical approaches.

The authors in [10] they established an advanced deep learning framework created to forecast Parkinson's disease using a dataset containing 42 voice recordings that have been subjected to pre-processing. Their study showcased improved accuracy compared to previous results. Nevertheless, it's noteworthy that the 81% accuracy achieved in 2018 is still regarded as relatively modest Within the scope of predicting Parkinson's disease.

The researchers Attained a 75% accuracy level and an 80% F1 score. In a research conducted by [20] they employed a 13-layer deep Convolutional Neural Network (CNN) model to detect Parkinson's disease from voice signals. Their experimentation involved a dataset comprising 20 patients. Despite achieving an 88% accuracy rate, their model made 361 incorrect predictions during the process.

In [21], introduced a novel classification approach for distinguishing between individuals with Parkinson's disease (PD) and those without, Utilizing dysphonia as a key factor, the researchers gathered data from 31 individuals, including 23 with PD and 8 healthy participants, who produced 195 sustained vowel sounds. Their method involved three crucial stages: feature computation, data preprocessing, feature selection, and the utilization of a linear kernel for classification. Impressively, the model achieved an accuracy of 91.4%.

In study [2], a novel deep learning framework, based on LSTM, is introduced to assess Parkinson's disease severity through gait pattern analysis. LSTM is utilized to capture temporal patterns in the data, eliminating the necessity for manual feature engineering and addressing the challenge of vanishing gradients. Furthermore, performance is enhanced through the incorporation of dropout, L2 regularization, and the application of the Adam optimizer. The results demonstrate outstanding performance with 98.6% accuracy in binary classification and 96.6% in multi-class classification, surpassing similar methods by 3.4%.

In [19] employed deep learning (DL) algorithms to detect Parkinson's disease (PD) and compared optimized and non-optimized methods. They found that K-fold cross-validation improved the accuracy of their approach. Another research [24] used artificial neural networks for PD detection.

In their study [23], researchers use machine learning algorithms on speech data to detect Parkinson's Disorder (PD) at an early stage, aiming to enable timely treatment and minimize its impact. They employ various models, including, Random Forest, Decision Tree , Naive Bayes, XGBoost, K Nearest Neighbor, and Support Vector Machine, along with Principal Component Analysis to reduce features. Their ensemble of the best-performing models achieves a 91% accuracy, improving disease recognition.

Our study investigates the application of advanced machine learning techniques, including XGBoost, Random Forest, MLP, and RNN, to enhance the classification and assessment of Parkinson's disease severity. Our dataset is meticulously curated, with features extracted from patient data. Our methodology encompasses a comprehensive data pre-processing phase that includes tasks such as data loading, managing missing values, and data normalization. Subsequently, we conduct exploratory data analysis, utilizing visual aids like histograms and box plots to gain insights into feature distribution and variability. These models undergo training and evaluation using performance metrics such as precision, recall, and F1-score. We enhance the effectiveness and ability to generalize of neural network models by integrating regularization methods such as L1 and L2.

Additionally, we utilize Principal Component Analysis (PCA) to simplify the intricacies of the dataset. The incorporation of GRU models enhances both the classification and severity assessment of PD. The integration of advanced ML methods provides insights into disease characterization. Notably, the analysis of severity percentages demonstrates the models' proficiency in evaluating severity. Remarkably, Random Forest achieves a severity analysis of 89.74%, with XGBoost closely following at 87.18%. These percentages bear clinical significance for well-informed medical decisions, emphasizing the importance of these models in Parkinson's disease analysis.

The results showcase promising achievements across all models, with XGBoost exhibiting an impressive precision of 97.4%. Random Forest, MLP, and RNN also demonstrate competitive precision rates. However, a more comprehensive analysis uncovers potential intricacies in accurately identifying negative instances. These insights underscore the significance of refining models to improve the detection of PD absence. Each model's performance is assessed using classification metrics, and detailed reports are generated to support in-depth comparisons.

The organization of this paper revolves around five distinct sections, ensuring a cohesive presentation of the content. Section 2 outlines the methodology employed in this study. Moving on to Section 3, we showcase the obtained results and engage in pertinent discussions. The conclusive insights drawn from this research are encapsulated in Section 4.

1. Methodology

1.1. Dataset

In this investigation, we make use of the dataset originally formulated by [15] with the principal objective of diagnosing PD by analyzing speech signals obtained from patients. The research encompasses voice samples collected from a total of 31 participants, among whom 23 exhibit Parkinson's disease while 8 are considered to be in a healthy condition. Within this dataset, there are 195 voice recordings, each characterized by 23 distinct attributes. The process of determining whether a particular voice recording pertains to an individual afflicted by Parkinson's disease or an individual in good health involves an assessment of 22 specific voice traits across the 195 biomedical voice recordings. This differentiation relies on the "status" column, where a value of 1 indicates the presence of Parkinson's disease and 0 indicates a healthy state. Further details concerning these 22 acoustic attributes can be found in table 1.

Fig. 1 depicts a histogram representing the acoustic properties of the dataset under analysis. Each chart shows the frequency of occurrences (on the y-axis) for different values (on the x-axis) associated with a particular attribute.

Table 1. Dataset description

Voice measure	Description
MDVP; Fo (Hz)	Average Vocal Fundamental Frequency
MDVP; Fhi (Hz	Maximum Vocal Fundamental Frequency
MDVP; Flo (Hz)	Minimum Vocal Fundamental Frequency
MDVP; Jitter (%)	Fundamental frequency perturbation (%)
MDVP; Jitter (Abs)	Absolute jitter in microseconds
MDVP; RAP	Relative Amplitude Perturbation
MDVP; PPQ	Five-point Period Perturbation Quotient
Jitter; DDP	Difference of differences between Cycles and period
MDVP; Shimmer	Shimmer Local amplitude perturbation
MDVP; Shimmer (dB)	Local amplitude perturbation (decibels)
Shimmer; APQ3	3-point Amplitude Perturbation Quotient
MDVP; APQ	11-point Amplitude Perturbation Quotient
Shimmer; APQ5	5-point Amplitude Perturbation Quotient
Shimmer; DDA	Average absolute difference between the amplitudes of consecutive periods
HNR, NHR	Harmonics-to-Noise Ratio, Noise-to-Harmonics Ratio
Status	Healthy (0) and Parkinson’s disease (1)
RPDE	Recurrence Period Density Entropy
DFA	Detrended fluctuation analysis
spread1, spread2, PPE	The fundamental frequency variation, pitch period entropy

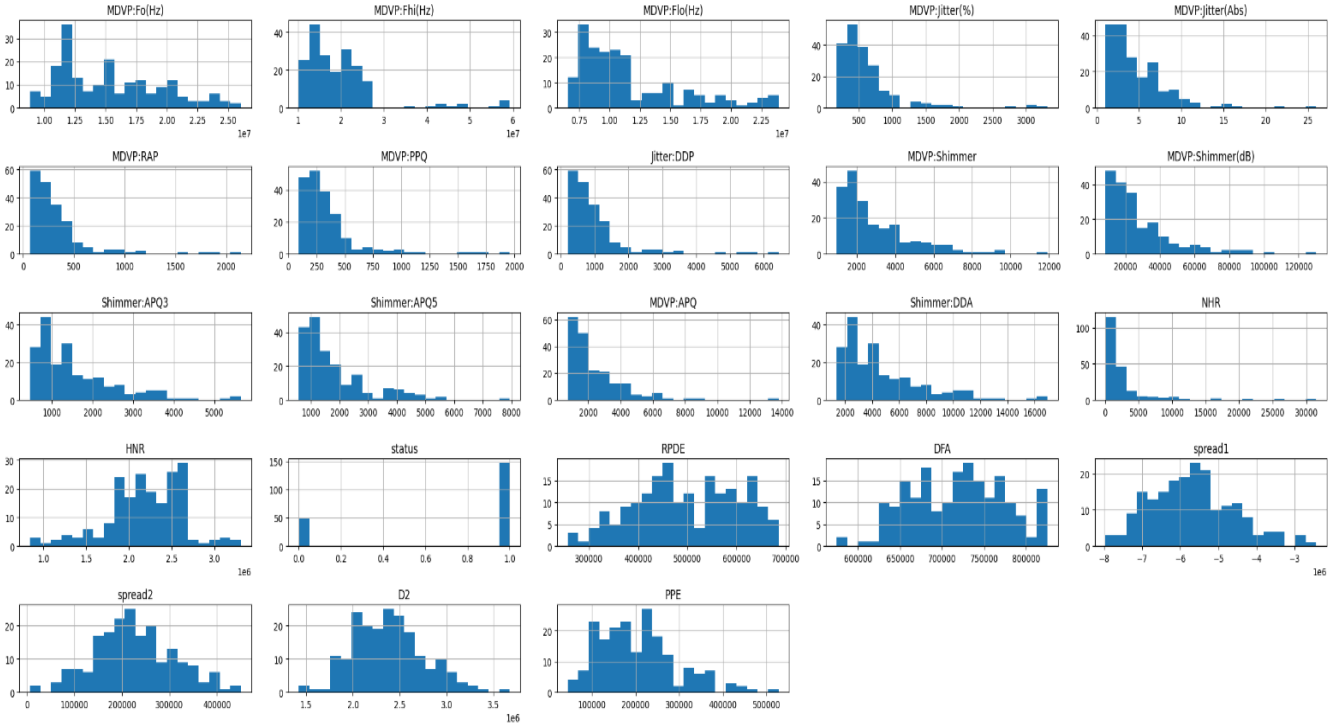


Fig. 1. Acoustic features histogram illustration

1.2. Feature extraction and advanced modeling techniques

This data analysis workflow commences by addressing missing data within the dataset. This process involves the removal of rows containing missing values and ensuring data conforms to the correct format by converting columns to numerical types while eliminating irrelevant columns. Following this, the dataset is partitioned into features (X) and the target variable (y). Subsequently, it is further divided into training and testing subsets. The data undergoes standardization using the "StandardScaler", and dimensionality reduction is executed through Principal Component Analysis (PCA).

Algorithm: Comprehensive Framework for PD Detection and Severity Assessment

Input: PD Dataset
Output: Disease Classification (PD or Healthy) and Severity Percentage

Commencement:
1. Load the PD dataset from a CSV file using Pandas.
2. Retrieve pertinent characteristics and transform them into an appropriate data format.
3. Preprocess the data:
   - Eliminate missing values through data cleansing.
   - Scale or standardize the features to the same range.
4. Split the dataset into training and testing subsets.
5. Perform PCA for dimensionality reduction:
   - Apply PCA to reduce feature dimensionality.

Training Phase:
6. Train a deep learning model:
   - Create a neural network with specified architecture.
   - Choose optimizer, activation functions, and regularization techniques.
   - Train the model on the training data using fit ().

Testing Phase:
7. Evaluate model performance on unseen test data:
   - Use the trained model to predict disease classification.
   - Compute metrics such as accuracy, precision, recall, and F1-score.
   - Create visual representations of the confusion matrix and ROC curve.
8. Calculate disease severity percentage using the trained model:
   - Utilize the trained model (GRU) to predict severity.
   - Apply the model to XGBoost, RandomForest, MLP, and RNN classifiers.
   - Calculate the percentage of severity based on different classifiers' predictions.

Conclusion.

When exploring models, a variety of neural network architectures are considered, commencing with a Sequential model that incorporates "Dense" and "Dropout" layers. This modeling approach utilizes the "Adam" optimizer and employs "sigmoid" activation for binary classification. Additionally, advanced ensemble techniques such as Random Forest are explored.

Furthermore, weight initialization in neural networks is fine-tuned, and advanced regularization methods, including L1 and L2 regularization, are applied to mitigate overfitting.

This comprehensive approach covers data pre-processing, feature extraction, model experimentation, and advanced techniques to ensure the development of robust and high-performing machine learning models.

The algorithm presents a thorough methodology crafted for the identification and evaluation of Parkinson's disease using voice signals. It commences by taking a Parkinson's disease dataset as input and systematically progresses through several stages. Initially, it divides the data, conducts sampling for training purposes, and then proceeds to train the model. During the training phase, a deep learning model is developed. Ultimately, the algorithm concludes by meticulously assessing the model's performance on unseen test data, providing valuable insights into its efficacy and suitability.

1.3. Data division

A traditional technique for splitting data entails a random division into two separate segments: one designated for training and the other for testing. The training subset is utilized to educate the model, while the test subset is employed to gauge its effectiveness. Typically, 80% of the data is assigned for training, leaving the remaining 20% for assessment. This methodology aligns well with your dataset.

1.4. Algorithms

XGBoost

XGBoost uses a convex loss function to quantify the difference between predicted and desired outcomes. This, combined with a penalty element to manage model complexity, seeks to minimize a regulated objective function involving L1 and L2 regularization. These functions essentially relate to regression trees. The iterative training process involves

integrating new trees that forecast residuals from prior trees. These additions are merged with existing trees to produce the final prediction. The term "gradient boosting" stems from employing gradient descent to minimize loss while incorporating new models [14]. Fig. 2 [11] shows the diagrammatic representation of XGBoost Algorithm.

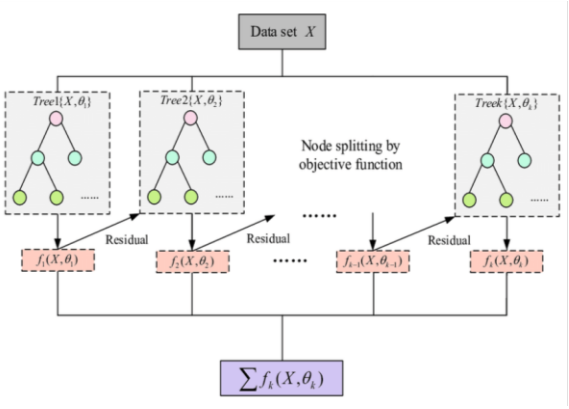


Fig. 2. Diagrammatic representation of XGBoost Algorithm

Random forest

A highly popular supervised machine learning algorithm adept at handling regression and classification tasks. It's based on ensemble learning, merging multiple classifiers to address complex challenges and bolster model performance. The "Random Forest" method employs multiple decision trees on diverse dataset subsets, combining their outcomes to boost dataset accuracy. Unlike relying on a single tree, it considers predictions from each tree to make its own based on majority consensus. Incorporating more trees heightens accuracy while guarding against overfitting [12].

MLP

The Multi-layer Perceptron, often referred to as MLP, represents a neural network architecture featuring densely connected layers that can adjust input dimensions to match desired output dimensions. It encompasses multiple layers of interconnected neurons, where the output of one neuron can serve as the input for another. An MLP typically comprises an input layer with one neuron for each input, an output layer with one neuron for each output, and it can include any number of hidden layers with varying node counts [6, 9]. In a standard MLP diagram, inputs are fed into the input layer, processed through the hidden layers, and ultimately yield outputs. All nodes shown in Fig. 3 within the MLP employ a sigmoid activation function to convert input values into a range spanning from 0 to 1 [4]. The sigmoid formula for this transformation is as follows: $\alpha(x) = 1 / (1 + \exp(-x))$.

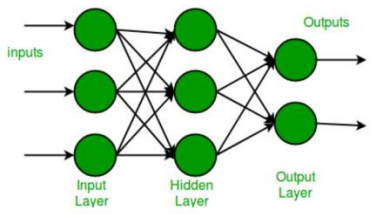


Fig. 3. A schematic diagram of a Multi-Layer Perceptron (MLP)

RNN

The network layers within an RNN establish cycles, which essentially involve utilizing the output of one layer as input for the next layer. Recurrent neural networks are commonly designed for tasks such as image description, automated translation, or processing natural language, as they assist in comprehending temporal or sequential data. RNNs find applications in tasks like automatic sleep apnea detection from nighttime ECG data [18] 16 and in the automated processing of speech [8].

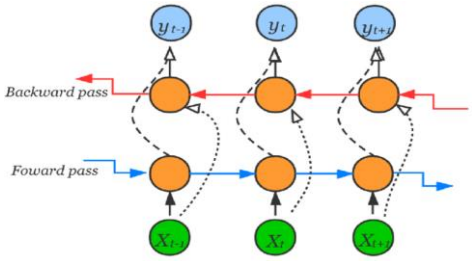


Fig. 4. The architecture of a recurrent network

2. Performance metrics

Performance metrics, as discussed [1], serve as essential tools for evaluating the effectiveness and precision of diverse models. These models rely on metrics such as accuracy, precision, recall, and the F1 score to make predictions based on provided data. In the assessment of these metrics, we consider four key components [16]:

- True Negatives (TN): in this context, we focus on correctly identifying negative cases.
- True Positives (TP): these represent instances where positive cases are correctly identified.
- False Positives (FP): conversely, false positives occur when the model incorrectly predicts a positive result for cases that are truly negative.
- False Negatives (FN): these arise when the model incorrectly predicts a negative outcome for cases that are actually positive.

Accuracy:

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN)$$

Precision:

$$\text{Precision} = TP / (TP+FP)$$

Recall:

$$\text{Recall} = TP / (TP+FN)$$

F1 score:

$$\text{F1Score} = (2\times\text{precision}\times\text{recall}) / (\text{precision} + \text{recall})$$

Specificity:

$$\text{Specificity} = TN / (TN+FP)$$

3. Results and discussion

3.1. Comparative analysis of machine learning models for classification performance

The role of performance comparison among models is to provide an objective and quantitative evaluation of their ability to minimize the classification task. Precision, measured individually for each model, helps quantify the accuracy of their positive predictions. In this specific case, we observe that XGBoost displays the highest precision, suggesting its capacity to minimize false positives. In comparison, the Random Forest exhibits slightly lower precision, indicating its ability to predict certain positive instances with slight uncertainty. MLP and RNN demonstrate similar precision performance, implying their capacity to maintain a good balance between true positives and false positives. This comparison sheds light on the respective strengths and weaknesses of the models, guiding the selection of the one that aligns best with the task's objectives and constraints.

The comparative analysis of machine learning model performances, as illustrated in table 2 for XGBoost, reveals significant differences. XGBoost stands out with the highest overall precision (0.974), showcasing its effectiveness in making accurate positive predictions. However, it is noteworthy that this high precision is not consistently maintained for the minority class, as highlighted in the table. While XGBoost excels in overall accuracy (97%), addressing concerns about its performance on the minority class could further enhance its effectiveness in diverse scenarios.

Table 2. Classification Report for XGBoost Model Performance

	precision	recall	F1-score	support
0	1.00	0.86	0.92	7
1	0.97	1.00	0.98	32
accuracy			0.97	39
macro avg	0.98	0.93	0.95	39
weighted avg	0.98	0.97	0.97	39

Table 3. Classification Report for Random Forest Model Performance

	precision	recall	F1-score	support
0	0.80	0.57	0.67	7
1	0.91	0.97	0.94	32
accuracy			0.90	39
macro avg	0.86	0.77	0.80	39
weighted avg	0.89	0.90	0.89	39

Table 4. Classification Report for MLP Model Performance

	precision	recall	F1-score	support
0	1.00	0.57	0.73	7
1	0.97	1.00	0.96	32
accuracy			0.92	39
macro avg	0.96	0.79	0.84	39
weighted avg	0.93	0.92	0.91	39

Table 5. Classification Report for RNN Model Performance

	precision	recall	F1-score	support
0	0.83	0.71	0.77	7
1	0.94	0.97	0.95	32
accuracy			0.92	39
macro avg	0.89	0.84	0.86	39
weighted avg	0.92	0.92	0.92	39

Turning to table 3, the Random Forest model exhibits slightly lower precision (0.897), indicating some challenges in maintaining precision, particularly for the minority class. This suggests that although Random Forest achieves an accuracy of 90%, further optimization may be beneficial for improving precision across both classes.

Meanwhile, as evident in tables 4 and 5, MLP and RNN demonstrate similar performances (0.923), with MLP showcasing a better balance between precision and recall. This balance is particularly crucial in scenarios where equal importance is placed on identifying both positive and negative instances. These findings emphasize the versatility of MLP in achieving a harmonious trade-off between precision and recall.

These insights underscore the need for a nuanced evaluation when determining the best-performing model. While XGBoost may excel in certain aspects, considering the broader context of precision, recall, and class-specific performance becomes imperative for making informed decisions. Additionally, exploring model-specific strengths and weaknesses, as outlined in the tables, allows for a more comprehensive understanding of how each model can be fine-tuned to achieve the best possible outcomes in diverse real-world applications.

In Fig. 5, the confusion matrices for different models offer a detailed look into the performance of the classification models. XGBoost exhibited a robust predictive ability, with only one misclassification out of 32 instances, showcasing its proficiency in accurately distinguishing between cases categorized as healthy (0) and unhealthy (1). The confusion matrices for MLP, RNN, and Random Forest models also displayed competitive outcomes, accurately predicting the majority of instances. However, a notable pattern emerged as these models encountered challenges in preventing false positives for the healthy class (0), as evident from the misclassifications.

These matrices provide a comprehensive understanding of the strengths and limitations of each model, streamlining their evaluation and selection for the classification task at hand, encompassing both healthy and unhealthy categories. By visually illustrating the actual and predicted class labels, the confusion matrices offer insights into the models' competence in correctly identifying instances from both classes. This information is crucial for informed decision-making about the appropriateness of each model for the classification task, aiding in the selection of the most suitable approach for the precise categorization of both healthy and unhealthy cases. The nuanced insights gained from the confusion matrices contribute to a more thorough assessment of each model's performance and guide the optimization process for enhancing their overall effectiveness in real-world applications.

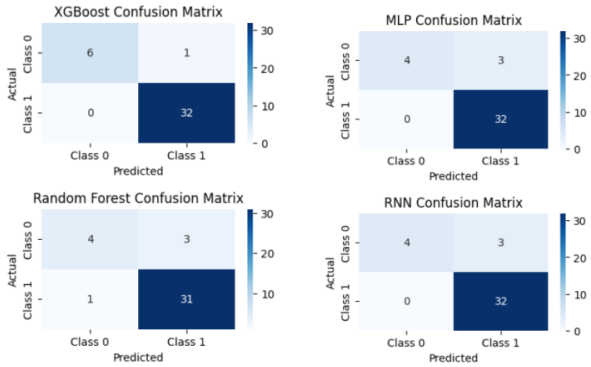


Fig. 5. Confusion Matrices for Different Models

3.2. Comparative Analysis of Machine Learning Models Using GRU for Disease Severity

The severity analysis percentages provide valuable insights into how effectively each machine learning model assesses the severity of the disease. Among the models examined, the XGBoost model achieved a severity analysis percentage of 87.18%, indicating its ability to accurately distinguish between severe and non-severe cases. The Random Forest model showed slightly better performance, with a severity analysis percentage of 89.74%. This suggests that the Random Forest model demonstrated a higher capability in correctly assessing the severity levels.

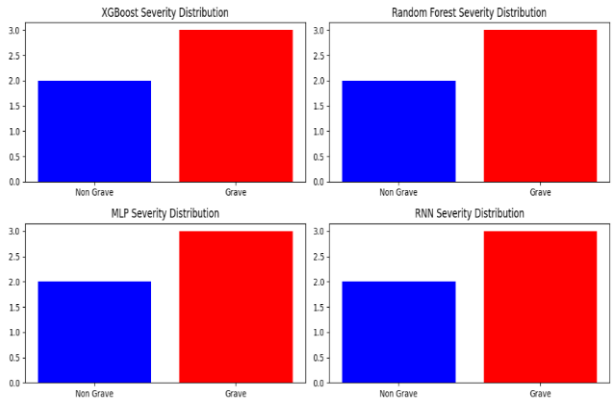


Fig. 6. Severity Distribution for Each Model

Additionally, in Fig. 6, the MLP and RNN models both exhibited a severity analysis percentage of 87.18%, aligning closely with XGBoost. This demonstrates that these models are consistent in their ability to analyze the severity levels, although they may face certain challenges, possibly in handling particular instances or class imbalances.

Regarding the GRU model, the provided code lacks its outcomes and severity analysis percentage. To comprehensively discuss its effectiveness, having the actual percentage calculated similarly to other models is crucial. Without this essential information, evaluating the GRU model's performance relative to others is challenging.

In summary, severity analysis percentages provide insight into each model's ability to evaluate disease severity. Notably, the Random Forest model stands out with the highest percentage. XGBoost, MLP, and RNN models consistently demonstrate their capabilities. To thoroughly evaluate the GRU model's efficiency, having its severity analysis percentage is essential for a comprehensive comparison.

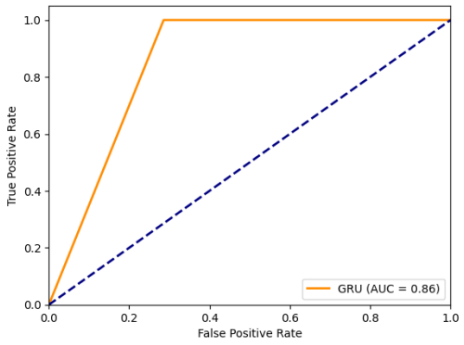


Fig. 7. Receiver Operating Characteristic (ROC) Curve for GRU Model

In Fig. 7, the achieved AUC of 0.86 for the GRU model demonstrates its effective performance in binary classification tasks. This robust performance signifies the model's reliability in distinguishing between positive and negative class samples, particularly in fields like healthcare, finance, and natural language processing. However, it's crucial to recognize that this performance should not be viewed in isolation. Future research endeavors should focus on enhancing model robustness, investigating the impact of class imbalances, and optimizing hyperparameters to maximize its utility in specific applications. In essence, this study contributes to advancing our understanding of GRU models in binary classification and paves the way for promising developments across various scientific and technological domains.

4. Conclusion

Our study harnessed advanced machine learning and deep learning techniques to classify Parkinson's disease and gauge its severity. Approaches like XGBoost, Random Forest, MLP, and RNN boosted disease categorization accuracy. Our work involved curated datasets, thorough preprocessing, and exploratory analysis to reveal crucial feature insights. Remarkably, our models achieved high precision, particularly XGBoost at 97.4%. However, challenges in identifying negative cases highlighted the need for model refinement. Furthermore, we introduced a GRU model for severity prediction, expanding our research horizon. By merging machine and deep learning, we contribute to disease diagnosis and prognosis advancement, facilitating potential personalized medical interventions. The findings emphasize computational methods' importance for timely interventions by healthcare professionals, ultimately enhancing the well-being of individuals grappling with this debilitating condition. In terms of future prospects, leveraging larger datasets and innovative approaches will likely enhance accuracy and personalized healthcare not only for Parkinson's disease but also for other neurological disorders.

References

[1] Abunadi I.: Deep and hybrid learning of MRI diagnosis for early detection of the progression stages in Alzheimer's disease. Connect. Sci. 34, 2022, 2395–2430.
[2] Balaji E. et al.: Automatic and non-invasive Parkinson's disease diagnosis and severity rating using LSTM network. Applied Soft Computing 108, 2021, 107463.
[3] Benba A., Jilbab A., Et Hammouch A.: Analysis of multiple types of voice recordings in cepstral domain using MFCC for discriminating between patients with Parkinson's disease and healthy people. International Journal of Speech Technology 19, 2016, 449-456.
[4] Bourdenx M. et al.: Identification of distinct pathological signatures induced by patient-derived ?-synuclein structures in nonhuman primates. Science advances 6(20), 2020, eaaz9165.
[5] Chaudhuri K. R., Schapira A. H.: Non-motor symptoms of Parkinson's disease: Dopaminergic pathophysiology and treatment. Lancet Neurol. 8, 2009, 464–474.
[6] El Bakali S., Ouadi H., Saad G.: Day-ahead seasonal solar radiation prediction, combining VMD and STACK algorithms. Clean Energy 7(4) (2023), 911–925.
[7] Erdogdu Sakar B., Serbes G., Sakar C. O.: Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. PLoS ONE 12(8), 2017, e0182428.
[8] Gelly G.: Reseaux de neurones recurrents pour le traitement automatique de la parole. Ph.D. thesis, Université Paris Saclay (COmUE), Paris 2017.
[9] Gheouany S. et al.: Experimental validation of multi-stage optimal energy management for a smart microgrid system under forecasting uncertainties. Energy Conversion and Management 291, 2023, 117309.
[10] Grover S., Bhartia S., Yadav A., Seeja K.: Predicting severity of Parkinson's disease using deep learning. Procedia Comput. Sci. 132, 2018, 1788-1794.
[11] Guo R. et al.: Degradation state recognition of piston pump based on ICEEMDAN and XGBoost. Applied Sciences 10(18), 2020, 6593.
[12] Gupta I. et al.: PCA-RF: an efficient Parkinson's disease prediction model based on random forest classification. 2022, arXiv preprint arXiv:2203.11287.
[13] Gürüler H.: A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method. Neural Computing & Applications 28(7), 2017, 1657-1666.
[14] Kumar A. et al.: A new Diagnosis using a Parkinson's Disease XGBoost and CNN-based classification model Using ML Techniques. International Conference on Advanced Computing Technologies and Applications – ICACTA. Coimbatore 2022, 1–6.
[15] Little M., McSharry P., Hunter E., Spielman J., Ramig L.: Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. Nat. Preced. 2008.
[16] Majdoubi O., Benba A., Hammouch A.: Classification of Parkinson's disease and other neurological disorders using voice features extraction and reduction techniques. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska – IAPGOS 13(3), 2023, 16-22.
[17] Poewe W., Seppi K., Tanner C., Halliday G., Brundin P., Volkmann J., Schrag A., Lang A.: Parkinson disease. Nat. Rev. Dis. Prim. 3, 2017, 17013.
[18] Prakash P., Sebban M., Habrard A., Barthelemy J.-C., Roche F., Pichot V.: Détection automatique des apnées du sommeil sur l'ECG nocturne par un apprentissage profond en réseau de neurones récurrents (RNN). Médecine du Sommeil 18(1), 2021, 43-44.
[19] Quan C., Ren K., Luo Z., Chen Z., Ling Y.: End-to-end deep learning approach for Parkinson's disease detection from speech signals. Biocybern. Biomed. Eng. 42, 2022, 556-574.
[20] Rehman A. et al.: Parkinson's disease detection using hybrid lstm-gru deep learning model. Electronics 12(13), 2023, 2856.
[21] Sharanyaa S., Renjith P. N., Ramesh K.: An exploration on feature extraction and classification techniques for dysphonic speech disorder in Parkinson's Disease. Inventive Communication and Computational Technologies – ICICCT. Singapore, 2022.
[22] Sriram T. V. S., Rao M. V., Narayana G. V. S., Kaladhar D. S. V. G. K.: Diagnosis of Parkinson disease using machine learning and data mining systems from voice dataset. 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications – FICTA. Berlin, 2014, 151–157.
[23] Tallapureddy G., Radha D.: Analysis of Ensemble of Machine Learning Algorithms for Detection of Parkinson's Disease. International Conference on Applied Artificial Intelligence and Computing – ICAAIC. Salem, 2022, 354–361.
[24] Yasar A., Saritas I., Sahman M., Cinar A.: Classification of Parkinson disease data with artificial neural networks. IOP Conf. Ser. Mater. Sci. Eng. 675, 2019, 012031.

M.Sc. Oumaima Majdoubi
e-mail: oumaima_majdoubi@um5.ac.ma

Received the Master's degree in Electrical Engineering from the National School of Arts and Crafts of Rabat, Mohammed V University, Rabat, Morocco, in 2022. She is a research student of sciences and Technology of the Engineer in Ecole Nationale Supériure d'Informatique et d'Analyse des Systémes ENSIAS. She is part of E2SN research team at ENSAM-RABAT. She is interested in the detection of neurological diseases based on artificial intelligence and signal processing.

http://orcid.org/0009-0000-2968-7975

Prof. Achraf Benba
e-mail: achraf.benba@ensam.um5.ac.ma

Received the Ph.D. degree in electrical engineering from ENSAM, Mohammed V University, Rabat, Morocco, in 2017. He is a member of E2SN teams at ENSAM, Mohammed V University. His interests are in speech processing for detecting people with neurological disorders and Cardiac.

http://orcid.org/0000-0001-7939-0790

Prof. Ahmed Hammouch
e-mail: ah.hammouch@gmail.com

Received the M.S. degree and the Ph.D. degree in automatic, electrical, electronic by the Haute Alsace University, Mulhouse, France, in 1993, and the Ph.D. degree in signal and image processing by the Mohammed V University, Rabat, Morocco, in 2004. From 1993 to 2013, He is a member of E2SN teams at ENSAM, Mohammed V University.

http://orcid.org/0009-0005-8691-6662

http://doi.org/10.35784/iapgos.5366 *received: 25.09.2023 | revised: 24.10.2023 | accepted: 27.11.2023 | available online: 20.12.2023*

AI EMPOWERED DIAGNOSIS OF PEMPHIGUS: A MACHINE LEARNING APPROACH FOR AUTOMATED SKIN LESION DETECTION

Mamun Ahmed, Salma Binta Islam, Aftab Uddin Alif, Mirajul Islam, Sabrina Motin Saima
Bangladesh Army International University of Science and Technology, Department of Computer Science and Technology, Comilla, Bangladesh

Abstract. *Pemphigus is a skin disease that can cause a serious damage to human skin. Pemphigus can result in other issues including painful patches and infected blisters, which can result in sepsis, weight loss, and starvation, all of which can be life-threatening, tooth decay and gum disease. Early prediction of Pemphigus may save us from fatal disease. Machine learning has the potential to offer a highly efficient approach for decision-making and precise forecasting. The healthcare sector is experiencing remarkable advancements through the utilization of machine learning techniques. Therefore, to identify Pemphigus using images, we suggested machine learning-based techniques. This proposed system uses a large dataset collected from various web sources to detect Pemphigus. Augmentation has been applied on our dataset using techniques such as zoom, flip, brightness, distortion, magnitude, height, width to enhance the breadth and variety of the dataset and improve model's performance. Five popular machine learning algorithms has been employed to train and evaluate model, these are K-Nearest Neighbor (referred to as KNN), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), and Convolutional Neural Network (CNN). Our outcome indicate that the CNN based model outperformed the other algorithms by achieving accuracy of 93% whereas LR, KNN, RF and DT achieved accuracies of 78%, 70%, 85% and 75% respectively*.

Keywords: pemphigus, blisters, augmentation, CNN

DIAGNOSTYKA PĘCHERZYCY Z WYKORZYSTANIEM SZTUCZNEJ INTELIGENCJI: PODEJŚCIE OPARTE NA UCZENIU MASZYNOWYM DO AUTOMATYCZNEGO WYKRYWANIA ZMIAN SKÓRNYCH

Streszczenie. *Pęcherzyca to choroba skóry, która może powodować poważne uszkodzenia ludzkiej skóry. Pęcherzyca może powodować inne problemy, w tym bolesne plamy i zakażone pęcherze, które mogą skutkować sepsą, utratą masy ciała i łaknienia, co może zagrażać życiu, próchnicą zębów i chorób dziąseł. Wczesne wykrycie pęcherzycy może uchronić przed śmiertelną chorobą. Uczenie maszynowe może zaoferować wysoce efektywne podejście do podejmowania decyzji i precyzyjnego prognozowania. Sektor opieki zdrowotnej doświadcza niezwykłych postępów dzięki wykorzystaniu technik uczenia maszynowego. Dlatego do identyfikacji pęcherzycy za pomocą obrazów zaproponowano techniki oparte na uczeniu maszynowym. Proponowany system wykorzystuje duży zbiór danych zebranych z różnych źródeł internetowych do wykrywania pęcherzycy. W zbiorze danych zastosowano augmentację przy użyciu technik takich jak powiększanie, odwracanie, zmiana jasności, zniekształcenie, zmiana wielkości, wysokość i szerokości, aby zwiększyć zakres i różnorodność zbioru danych oraz poprawić wydajność modelu. Do uczenia i oceny modelu wykorzystano pięć popularnych algorytmów uczenia maszynowego, są to: K-Nearest Neighbor (określany jako KNN), drzewo decyzyjne (DT), regresja logistyczna (LR), las losowy (RF) i konwolucyjną sieć neuronowa (CNN). Uzyskane wyniki wskazują, że model oparty na CNN był lepszy od innych algorytmów, osiągając dokładność na poziomie 93%, podczas gdy LR, KNN, RF i DT osiągnęły dokładność odpowiednio 78%, 70%, 85% i 75%.*

Słowa kluczowe: pęcherzyca, pęcherze, augmentacja, CNN

Introduction

Normal and healthy skins are like a blessing to human as it works like a shield for human body. Yet the world is covered by many skin diseases. Nowadays, millions of people worldwide suffer from dermatological conditions every year, which constitute a serious threat to health. In our research, our main focus is on Pemphigus. Pemphigus is a rare autoimmune disorder characterized by the immune system's erroneous targeting of healthy skin and mucous membrane cells, causing painful blisters and sores. Immune system mistakenly targeting the proteins that hold skin cells together is the root cause of it, leading to the separation of skin layers and the formation of blisters. Pemphigus can affect anyone. The age range between 40 and 60 can frequently be the one most prone to its effects [15]. According to epidemiological research conducted in several European countries, pemphigus appears to be more uncommon at higher latitudes than in lower latitudes [18]. Pemphigus encompasses various subtypes, such as pemphigus foliaceus, pemphigus vegetans, and pemphigus vulgaris, which differ in their severity, location of blistering, and associated symptoms. Pemphigus has an elusive exact cause. According to research, both environmental and genetic factors may affect how you are diagnosed. Rarely, the condition can be brought on by specific drugs prescribed to treat certain illnesses. According to some research, certain HLA genes, which are immune system-building genes, predispose you to developing particular kinds of the disease. Treatment typically involves immunosuppressive medications to control the autoimmune response and promote healing of the skin and mucous membranes. Pemphigus can be a severe and lethal condition, especially if left untreated. Traditional diagnosis of pemphigus relies on evaluating the patient's clinical symptoms, performing a histopathological examination, and conducting direct immunofluorescence testing. In most cases people are not aware of stages of bare diseases or its type. Certain dermatological conditions may remain asymptomatic for an extended period, enabling the diseases to progress and disseminate unnoticed. Dermatologists can have trouble correctly identifying skin conditions, necessitating pricey laboratory testing to determine the specific type and stage of the problem. Advancements in medical technology utilizing photonics and lasers have facilitated rapid and precise diagnosis of skin diseases. Nevertheless, the expense associated with such diagnostic methods remains a significant barrier, making them unaffordable for many individuals [7]. Consequently, we propose utilizing image processing techniques for skin disease diagnosis. Basically, our primary aim is to detect the Pemphigus. The machine learning techniques is used to perform this task. Computers may learn from large datasets using machine learning, a type of artificial intelligence (AI), which makes use of algorithms and statistical models, empowering them to make predictions and informed decisions without the need for explicit programming for every potential scenario.

1. Literature review

As of today, no papers have been published on detecting pemphigus using machine learning based approach. Various skin diseases are the focus of the majority of reviewed papers.

The article authored by Jainesh Rathod and colleagues, introduces an image-based machine learning system that automatically detects skin disorders [11]. The suggested method uses softmax as the classifier in a convolutional neural network (CNN) architecture. The authors have also developed a web application based on this model. According to their evaluation results, the proposed method achieved an accuracy of approximately 70%.

In the scholarly article authored by R. Sumithra, M. Suhilb, D. S. Guruc et al. [13], in this paper, the authors propose an innovative methodology for skin disease segmentation and classification, employing two distinct machine learning algorithms: k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM) to achieve this task [13]. The results of the experiments are promising, with F-measures of 34% and 46.71% achieved by SVM and k-NN classifiers, respectively. Additionally, the fusion of k-NN and SVM classifiers achieved an even higher F-measure of 61%.

The scientific article discusses research that looked at 98 pemphigus vulgaris patients' long-term results after receiving oral treatment in north-west Italy [1]. In this study, the authors employed a Logistic Regression (LR) algorithm to analyse the data, which resulted in an accuracy rate of 84.21%.

An approach for grouping skin images utilizing navigation (navi) for categorization is suggested in the research paper written by Polap D. et al. [10]. The authors outline a smart home system that analyzes occupants' skin health using integrated sensors and artificial intelligence methods. The proposed method incorporates the Scale-Invariant Feature Transform (SIFT) technique to detect crucial regions in the images. Subsequently, the classification and segmentation tasks are accomplished using the Support Vector Machine (SVM) and Convolutional Neural Network (CNN) algorithms, resulting in accuracy and precision rates of 82% and 84% respectively.

According to a study by Tanzina Afroz Rimi et al. [12], the CNN model was proposed to identify numerous skin conditions. The objective of this research was to develop a prototype for a neural network-based system capable of identifying skin disorders. The authors selected the convolutional neural network (CNN) as their preferred neural network for the task. The accuracy attained was approximately 73%.

The scholarly paper by Dr. T. Kameswara Rao, P. Chamanthi, N. Tharun Kumar, R. Lakshmi Amulya, and M. Uday Sagar et al. [6] uses Convolutional Neural Networks (CNN) and an ensemble model that includes the Inception, DenseNet, and VGG16, architectures to present a unique method for skin disease identification. The accurate detection of a variety of skin disorders, including basal cell carcinoma, actinic keratoses, melanoma, vascular lesions, benign keratoses, dermatofibromas, and melanocytic nevi is the primary goal of the proposed system The results of the study show that the accuracy of the CNN model is between 71% and 75%. The VGG16, DenseNet, and Inception architectures achieve accuracy rates of 80.3%, 82.3%, and 80.4%, respectively. The ensemble model, which combines all three architectures, achieves the highest accuracy of 83% to 85%. This research provides a promising avenue for further development of intelligent skin disease detection systems that can improve patient care and diagnosis.

The scientific article authored by Shuchi Bhadula, Sachin Sharma, Piyush Juyal, Chitransh Kulshresth et al. [2] provides an extensive review of the use of machine learning algorithms for precise identification and classification of various skin disorders. To determine the exact type of skin diseases, the scientists investigated five different algorithms, including "Kernel SVM", "Logistic Regression", "Random Forest", "CNN" and "Naive Bayes". The studies were run on a large dataset of skin image data, and the results showed that Random Forest, Logistic Regression, and Naive Bayes all performed better than Logistic Regression in terms of accuracy (49%, 73.36% and 73.76%, respectively) [2]. Insights from this study investigating the viability and effectiveness of machine learning algorithms for skin disease diagnosis might be used to create future healthcare systems that are more advanced and intelligent. A comprehensive summary is shown in table 1.

Table 1. Comprehensive summary of machine learning techniques used in reviewed papers

Ref:	Algorithm
"Diagnosis of skin diseases using Convolutional Neural Networks" [11]	CNN
"Segmentation and classification of skin lesions for disease diagnosis" [13]	KNN, SVM
"Long-term evaluation of pemphigus vulgaris: a retrospective consideration of 98 patients treated in an oral medicine unit in North-West Italy" [1]	LR
"An Intelligent System for Monitoring Skin diseases" [10]	CNN, SVM
"Derm-NN: Skin Diseases Detection Using Convolutional Neural Network" [12]	CNN
"Skin Disease Detection Using Machine Learning" [6]	CNN, VGG-16, DenseNet, Inception
"Machine Learning Algorithms based Skin Disease Detection" [2]	LR, CNN, RF, Naïve Bayes

[@] – indicate the reference papers respectively

2. Machine learning

Machine learning is used in a wide range of fields, including the creation of autonomous cars as well as activities like speech and image recognition, natural language processing, and predictive analytics. Three primary categories of machine learning algorithms exist [9]:

Supervised: in the machine learning paradigm known as "supervised machine learning", a computer program learns from labeled data to make predictions or judgments based on incoming data. Making a model that can predict output values for unknown inputs with accuracy is the goal. Classification and regression are the two primary categories for supervised learning algorithms. These algorithms find applications in diverse fields, including image classification, speech recognition, and medical diagnosis.

Unsupervised: unsupervised machine learning involves training a computer program on an unlabeled dataset to uncover patterns or relationships within the data without prior knowledge of expected outcomes. The primary goal is to discover hidden structures or groupings within the data. Clustering and dimensionality reduction are the two basic forms of unsupervised learning methods. These algorithms find application in diverse domains, including fault detection, market segmentation, and image compression, among others.

Reinforced: through a system of incentives and penalties, the machine learning technique called reinforcement learning teaches a computer program to make judgments. The objective is to maximize the total reward accumulated over time. This method involves teaching an agent to act in a given environment, getting feedback in the form of incentives or punishments, and then changing its behavior in response. Reinforcement learning finds application in diverse domains, such as game playing, robotics, and autonomous vehicles, among other areas.

3. Problem identification

Upon reviewing the aforementioned reference papers, it appears that none of them have focused specifically on pemphigus, a potentially life-threatening autoimmune skin disease. On the contrary, studies have investigated a range of different dermatological conditions, including but not limited to benign keratosis, melanoma, actinic keratoses, bullae, vascular lesions, squamous cell carcinoma, shingles, seborrheic keratosis, basal cell carcinoma, Stevens-Johnson syndrome (SJS), dermatofibroma, acne, lichen planus, eczema, melanocytic nevi, and common skin infections. Different machine learning algorithms have been applied in these studies, such as Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Random Forest, K-nearest Neighbor (KNN), Naive Bayes, Logistic Regression, VGG-16, DenseNet, and Inception, to accomplish skin disease detection and classification with varying degrees of accuracy. However, the absence of research on pemphigus indicates a gap in the current knowledge, highlighting the need for further studies to develop effective diagnostic and treatment approaches for this rare and complex skin condition.

4. Our proposed work

The aim of our work is to detect Pemphigus using traditional system model. We applied five well-known machine learning strategies, namely Logistic regression (LR), K-nearest neighbor (KNN), Random Forest (RF), Decision tree (DT), and Convolutional Neural Network (CNN).

With the help of the Sklearn libraries, Pandas, Matplotlib, and other required libraries, the suggested work has been put into practice. The data were gathered from web sources [14, 16, 17, 19, 20] and incorporated in our own dataset. The dataset consists of true and false data of pemphigus.

5. Dataset description

5.1. Data collection

In any kind of research dataset is must needed. Because, a dataset can be considered as a primary part of any research work. But, for our work we couldn't get any readymade dataset from any particular online sources like Kaggle, Google Colab etc. So, the dataset we used in our research consists of approximately 600 data gathered from various online sources. This dataset is entirely made by ourselves.

5.2. Data organize

We considered these 600 data to be our initial dataset. The initial dataset is divided into two subsets: one of them are Pemphigus Vulgaris image collections where there are approximately 158 images are here. And the second one is the collection of 442 images that we named others. This subset contains a lot more different kinds of images of skin diseases and normal skin.

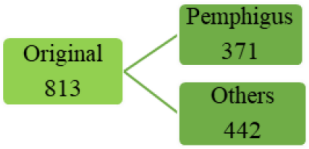


Fig. 1. Visual representation of data distribution in different categories

Fig. 2. Sample image (original): a) pemphigus, b) pemphigus, c) other skin diseases, d) pemphigus, e) normal skin, f) pemphigus, g) pemphigus, h) other skin diseases, i) pemphigus

5.3. Data augmentation

As our initial dataset got failed to provide a good accuracy in our implementation, we have to augment this initial dataset to generate an adequate number of images to get a high accuracy. This portion contains about 3500 images. This augmented dataset is a combination of two subsets. One is Pemphigus and another is others. Both of these subsets contain 1367 and 2133 images respectively.

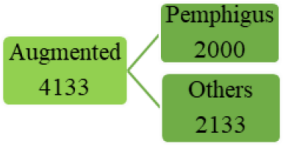


Fig. 3. Visual representation of augmented data distribution in different categories

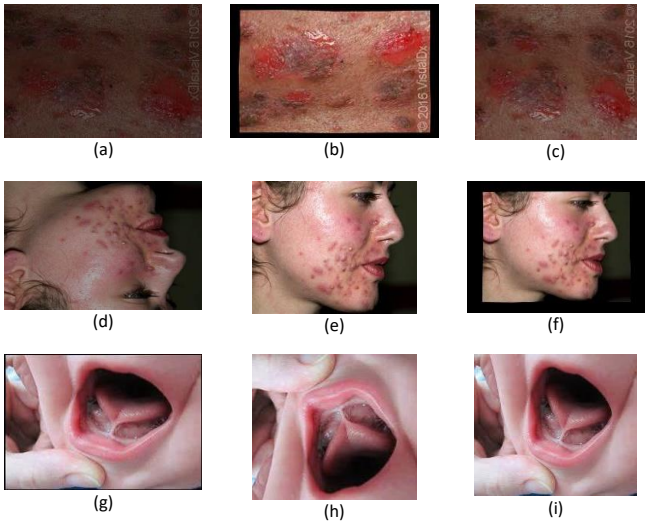


Fig. 4. Sample image (after augmentation): a) pemphigus (zoom), b) pemphigus (brightness), c) pemphigus (flip), d) other skin diseases (rotation), e) other skin diseases (zoom), f) other skin diseases (distortion), g) normal skin (zoom), h) normal skin (flip), i) normal skin (brightness)

5.4. Data preprocessing

We needed to solve a few issues that come up when importing the data in order to get a high performance of pemphigus vulgaris identification and prediction. such as image size and colour contrast [3]. Before uploading them to the server for processing, the image resizer program in Python automatically resizes every image. Therefore, the primary goal of this stage is to eliminate background noises from photographs of pemphigus vulgaris and other images.

6. Dataset splitting

The technique of splitting a dataset into distinct subsets is known as data splitting in machine learning. This procedure is generally used for training and testing machine learning models. Data splitting is used to check the model's performance on new data to prevent overfitting. The dataset is often separated into three subsets: test, validation, and training. The validation set is used to fine-tune the hyperparameters of the model, whereas the training set is used to develop the model. Lastly, to evaluate the model's performance on new, unseen data that was not used during its development, we utilized a test set. In this method, we partitioned the dataset, allocating 60% for the training set, 20% for the validation set, and the remaining 20% for the test set.

7. Methodology

Our research is mainly based on only one stage of Pemphigus Vulgaris prediction. The conventional Pemphigus Vulgaris disease prediction system was previously presented without incorporating a hyper parameter tuning method for the Machine Learning algorithms. The corresponding block diagram is illustrated in figure 5.

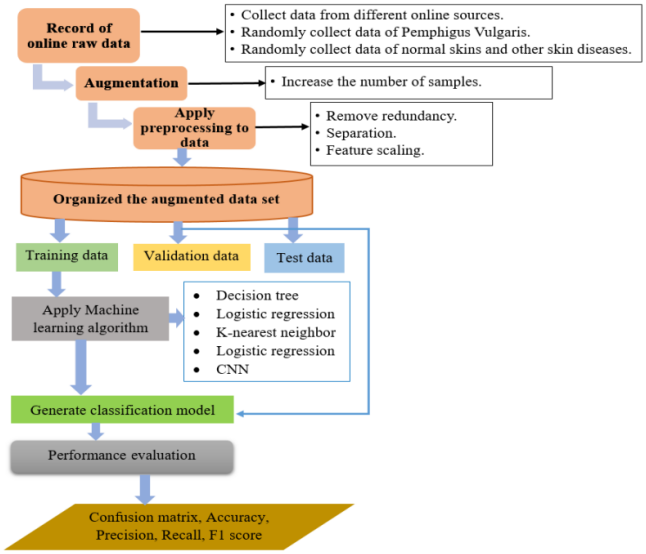


Fig. 5. Unveiling the work process: a step-by-step methodology

8. Model generation

During this phase, machine learning algorithms are employed to train various classification models using the training dataset. Individual samples are categorized based on these created models after the model development. In this particular scenario, a conventional approach utilizes five machine learning algorithms, namely LR, KNN, RF, DT, and CNN classifiers, to construct the models. Subsequently, the test set is classified using five models, and their performance is evaluated. Notably, the traditional system does not incorporate any hyper parameter tuning methodology and relies on default parameter settings for generating these five classification models [5].

9. Machine learning algorithms

In the model generation process, Logistic regression (LR), K-nearest neighbor (KNN), Random Forest (RF), Decision tree (DT), and Convolutional Neural Network (CNN) classifiers are used as machine learning algorithms. In our Pemphigus Vulgaris prediction system these five machine learning algorithms are used. Because they outperform other machine learning algorithms.

Logistic regression is a supervised ML technique that classifies observations based on independent variables. The logistic function is utilized in this context to compute the probability of an observation being classified into a specific class.

The non-parametric machine learning technique K-nearest neighbor (KNN) is used for both classification and regression problems. To determine the result of a specific observation, K nearest neighbors are chosen from the training set.

An ensemble of decision trees is used in the supervised machine learning method known as random forest to do both classification and regression problems. It mitigates overfitting by combining multiple decision trees and employing subsets of input variables and training data.

A simple supervised machine learning technique called a decision tree is utilized for both classification and regression problems. It divides data into subsets during training and builds a tree model to predict new observations' class or numerical value based on input attributes.

Convolutional Neural Networks (CNNs) are used for computer vision tasks such as object identification and recognition. They mimic the operation of the human visual cortex and consist of convolutional, pooling, and fully linked layers. CNNs have advantages over typical ML methods, such as handling spatial correlations and automatically learning features from data.

10.Performance evaluation

At this point, the confusion matrix is used by the algorithm to assess how well the training and test sets performed. The generated confusion matrix is then utilized to compute and analyze various performance metrics, including accuracy, precision, recall, and F1 score, for both models. The accuracy ratio is the proportion of properly recognized observations to all observations, whereas precision is calculated by dividing the total number of anticipated positive samples by the percentage of successfully categorized positive samples. By dividing the total number of positively categorized samples by the total number of positively tested samples, recall is calculated. The mathematical expressions for accuracy, precision, recall, and F1 score are depicted by equations (1), (2), (3), and (4) correspondingly [4].

Accuracy, also referred to as classification rate, is the percentage of correctly predicted overall results. The skin condition is correctly predicted if the precision is higher. Accuracy can be calculated as follows:

$$Accuracy = \frac{TP + TN}{TFP + TN + FP + FN} \tag{1}$$

The ratio of accurately anticipated positive cases to all of the actual positive instances in the test data is called recall, also known as sensitivity or true positive rate, also known as sensitivity. pemphigus disease is correctly categorized according to Higher Recall. The relationship given below explains the recall:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

A measure of precision compares the percentage of accurately anticipated positive observations to the total number of positive observations. It provides an assessment of the accuracy of positive predictions made by a model. The relationship given below provides precision:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

In order to provide a fair evaluation of a model's performance, the F1 score is a statistic that combines recall and accuracy into a single value [8]. It is determined by calculating the harmonic mean of recall and accuracy, ranging between 0 and 1, with 1 representing the best possible score. The formula for computing the F1 score is as follows:

$$F1\ score = \frac{2 \times (Recall \times Precision)}{Recall + Precision} \tag{4}$$

Here: TP – true positive, TN – true negative, FP – false positive, FN – false negative.

11.Result analysis

In our research, the machine learning algorithms that we used has default parameters. The overall performance of different algorithms and the comparison between these algorithms are given below in a tabular format:

First of all, we have used the Logistic Regression (LR) which is fitted and executed with parameters of $C = 10^5$ and max iteration = 1000 and we have got an accuracy of 76%, precision of 75%, recall of 76% and F1 score of 75%.

Now it comes KNN algorithm where the algorithm is fitted and executed the model with the parameters of no of neighbour = 50 and weights=distance and found 61%, 73%, 67% and 59% of accuracy, precision, recall and F1 score respectively.

In this period, we have worked with random Forest (RF) algorithm where the algorithm is fitted and executed the model with the parameters of maximum depth = 100 and random state = 0 and found 87%, 88%, 85% and 86% of accuracy, precision, recall and F1 score respectively.

In the fourth phase, the Decision Tree (DT) model is fitted and run using the best splitter and Gini index parameters, and the results are 100% accuracy, 100% precision, 100% recall, and 100% F1 score, respectively. This DT model predicts the test set and yields accuracy, precision, recall, and F1 score of 73%, 72%, 73%, and 73%, respectively.

In the last phase the CNN model is fitted and executed the model with the parameters of optimizer = 'adam', three layer and epochs=30 and found the accuracy 93%. Where precision, recall, and F1 score of 54%, 54.5% and 54% respectively. Performance evaluation is shown in table 2.

Table 2. Performance evaluation of classification models

Traditional System		Performance Evaluation			
Machine Learning Algorithms	Parameters	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Logistic Regression	C = 1 Max_iter = 1750	78%	77%	77%	77%
K-nearest Neighbor	No of neighbor = 5 Weights = 'uniform'	70%	73%	73%	70%
Random Forest	max_depth = 100 random_state = 0	85%	87%	83%	84%
Decision Tree	Criterion = 'gini' Min samples leaf = 1 Min samples split = 2 Spliter = 'best'	75%	74%	75%	74%
CNN	Optimizer = 'adam' Layer = 3 Epochs = 30	93%	54%	54.5%	54%

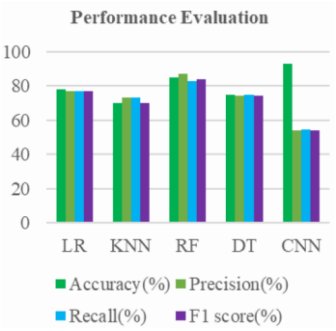


Fig. 6. Comparative graphical representation of performance evaluation of different algorithms

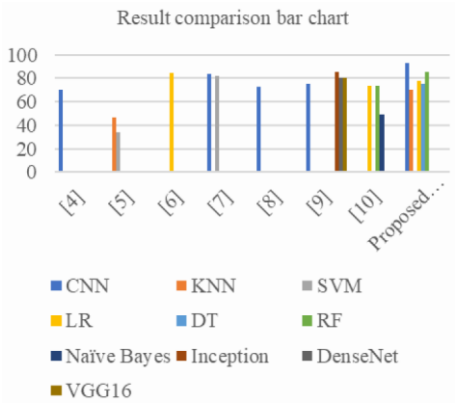


Fig. 7. Comparative analysis of machine learning algorithms in detecting other skin diseases and pemphigus

By leveraging the insights gained from these studies, we were able to draw meaningful conclusions about the potential of machine learning for pemphigus detection. While there is still much to be explored in this area, our study contributes to the broader understanding of the potential of machine learning in medical diagnosis and disease detection.

As noted previously, there is currently no published research specifically focused on the application of machine learning algorithms to detect pemphigus. However, we conducted a comparative analysis by reviewing existing literature on the use of machine learning for skin disease detection. Our analysis included a thorough examination of the methods and results of each study.

12.Result visualization

This section aims to provide a comprehensive visualization of the results obtained from applying our machine learning algorithm to the dataset. By presenting these visualizations, we seek to offer a deeper understanding of the algorithm's performance, behavior, and effectiveness. Our results demonstrate that the proposed machine learning-based system can effectively detect pemphigus from skin lesion images, with promising results for future improvements and clinical applications.

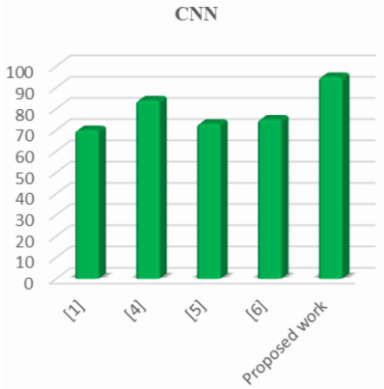


Fig. 8. Comparative graphical representation of performance evaluation of CNN in detecting other skin diseases and pemphigus

Fig. 9. Some predicted results of our proposed model

13. Conclusion

The field of medical science is experiencing a revolutionary transformation with the advent of machine learning, as it has promising potential to help demystify. An image detection system has been proposed that is capable of effectively identifying skin lesions caused by the rare autoimmune disorder known as pemphigus. Five algorithms have been applied on augmented image for predicting this disorder. Among these five algorithms CNN provide the best accuracy. In the near future, we plan to incorporate various types of classifiers to enhance the detection rate of our proposed system. As it is an infrequent autoimmune disease, the source of collecting dataset is inherently limited. To improve the detection rate it may be beneficial to enrich the dataset and diversifying the algorithms. By doing so, we may be able to uncover more-subtle patterns and features that can expand model's capabilities and lead to better generalization of the models. Overall, our study highlights the significant potential of machine learning in the field of dermatology.

Reference

[1] Arduino P. G. et al.: Long-term evaluation of pemphigus vulgaris: a retrospective consideration of 98 patients treated in an oral medicine unit in north-west Italy. Journal of Oral Pathology & Medicine 48(5), 2019, 406–412 [http://doi.org/10.1111/jop.12847].

[2] Bhadula S. et al.: Machine Learning Algorithms based Skin Disease Detection. International Journal of Innovative Technology and Exploring Engineering – IJITEE 9(2), 2019, 4044-4049 [http://doi.org/10.35940/ijitee.B7686.129219].

[3] Elngar A. A.: Intelligent System for Skin Disease Prediction using Machine Learning. 3rd International Conference on Smart and Intelligent Learning for Information Optimization 1998, 2021 [http://doi.org/10.1088/1742-6596/1998/1/012037].

[4] Hashi E. K., Md. Shahid Uz Zaman: Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction. Journal of Applied Science & Process Engineering 7(2), 2020, 631–647 [http://doi.org/10.33736/jaspe.2639.2020].

[5] Jonnavithula S. K. et al.: Role of Machine Learning Algorithms Over Heart Diseases Prediction. 2nd International Conference on Sustainable Manufacturing, Materials and Technologies 2292(1), 2020, [http://doi.org/10.1063/5.0030743].

[6] Kameswara Rao T. et al.: Skin Disease Detection Using Machine Learning. UGC CARE Listed (Group-I) Journal 11(12), 2022, 1593–1604 [http://doi.org/10.48047/IJFANS/V11/I12/171].

[7] Kumar A., Shetty P., Balipa M., Rao B., Puneeth B., Shravya: An efficient technique to detect skin Disease Using Image Processing. International Conference on Artificial Intelligence and Data Engineering – AIDE, Karkala 2022, 35–40 [http://doi.org/10.1109/AIDE57180.2022.10060001].

[8] Kumar V. B., Kumar S. S., Saboo V.: Dermatological disease detection using image processing and machine learning. Third International Conference on Artificial Intelligence and Pattern Recognition – AIPR, Lodz, 2016, 1–6 [http://doi.org/10.1109/ICAIPR.2016.7585217].

[9] Mahesh B.: Machine Learning Algorithms – A Review. International Journal of Science and Research – IJSR 9(1), 2020, 381–386.

[10] Połap D. et al.: An Intelligent System for Monitoring Skin diseases. Special Issue From Sensors to Ambient Intelligence for Health and Social, 2018 [http://doi.org/10.3390/s18082552].

[11] Rathod J. et al.: Diagnosis of skin diseases using Convolutional Neural Networks. Second International Conference on Electronics, Communication and Aerospace Technology – ICECA, Coimbatore, 2018, 1048–1051 [http://doi.org/10.1109/ICECA.2018.8474593].

[12] Rimi T. A. et al.: Derm-NN: Skin Diseases Detection Using Convolutional Neural Network. 4th International Conference on Intelligent Computing and Control Systems – ICICCS. Madurai, 2020, 1205–1209 [http://doi.org/10.1109/ICICCS48265.2020.9120925].

[13] Sumithra R., Suhilb M., Guruc D. S.: Segmentation and classification of skin lesions for disease diagnosis. Procedia Computer Science 45, 2015, 76–85 [http://doi.org/10.1016/j.procs.2015.03.090].

[14] American Academy of Dermatology. https://www.aad.org/public/diseases/a-z/pemphigus-symptoms (accessed: 21.01.2023).

[15] Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/21130-pemphigus (accessed: 20.04.2023)

[16] DermNet. https://dermnetnz.org/images/pemphigus-vulgaris-images (accessed: 10.01.2023).

[17] DermNet. https://dermnetnz.org/topics/pemphigus-foliaceus (accessed: 07.02.2023).

[18] National library of medicine. https://www.ncbi.nlm.nih.gov/books/NBK560860/ (accessed: 29.04.2023)

[19] NHS. https://www.nhs.uk/conditions/pemphigus-vulgaris/ (accessed: 03.01.2023).

[20] WebPathology. https://www.webpathology.com/image.asp?n=2&Case=697 (accessed: 20.02.2023).

M.Sc. Mamun Ahmed
e-mail: mamun.cse@baiust.ac.bd

Mamun Ahmed holds an M.Sc. in Signal Processing from Blekinge Institute of Technology (BTH), Sweden, and a Bachelor's degree in CSE from CUET, Bangladesh. He served as an RF Engineer at Motorola, Bangladesh, from 2006 to 2009. Since 2013, he has been an esteemed faculty member (assistant professor) in the Department of Computer Science and Engineering at Bangladesh Army International University of Science and Technology (BAIUST), Cumilla, Bangladesh. Published about 23 scientific papers.

http://orcid.org/0000-0002-3980-3981

Salma Binta Islam
e-mail: salma.islam3099@gmail.com

Recent graduate from Department of Computer Science and Engineering from Bangladesh Army International University of Science and Technology. Research interest: machine learning in medical science, computer vision.

http://orcid.org/0009-0004-9975-4861

MD. Aftab Uddin Alif
e-mail: alifbaiust@gmail.com

MD. Aftab Uddin Alif complete his Bachelor's degree with First-Class First in Computer Science and Engineering (CSE) from Bangladesh Army International University of Science and Technology (BAIUST), Bangladesh. He brings his expertise and knowledge to the field, having served as an Assistant Engineer at IDS Tech Solution in Bangladesh since July 2023. MD. Aftab Uddin Alif's academic achievements and professional experience position him as a valuable contributor in the field of computer science and engineering.

http://orcid.org/0009-0001-8461-1129

MD. Mirajul Islam
e-mail: mirajulislam76779@gmail.com

MD. Mirajul Islam Ratul is a dedicated researcher in the field of Computer Science and Engineering (CSE). He completed his Bachelor of Science (B.Sc.) in CSE from the prestigious Bangladesh Army International University of Science & Technology, where he exhibited a strong passion for academic excellence and research. His research focus has been on the cutting-edge application of machine learning techniques in the field of medical diagnostics. In particular, Ratul's work has centered around the development of novel approaches for the early detection of pemphigus, a rare and potentially debilitating autoimmune skin disease.

http://orcid.org/0009-0006-1215-1422

Sabrina Motin Saima
e-mail: smsaima2000@gmail.com

Sabrina Motin Saima achieved a First-Class in Computer Science and Engineering (CSE) for her Bachelor's degree from Bangladesh Army International University of Science and Technology (BAIUST), Bangladesh. With prior experience as an IT Support intern at Electricity Generation Company of Bangladesh Ltd., Sabrina Motin Saima brings her expertise and knowledge to the field, establishing herself as a valuable contributor in the realm of computer science and engineering.

http://orcid.org/0009-0005-7319-7259

http://doi.org/10.35784/iapgos.4464 *received: 31.08.2023 | revised: 19.10.2023 | accepted: 27.11.2023 | available online: 20.12.2023*

OPTIMIZING ULTRASOUND IMAGE CLASSIFICATION THROUGH TRANSFER LEARNING: FINE-TUNING STRATEGIES AND CLASSIFIER IMPACT ON PRE-TRAINED INNER-LAYERS

Mohamed Bal-Ghaoui, My Hachem El Yousfi Alaoui, Abdelilah Jilbab, Abdennaser Bourouhou
Mohammed V University in Rabat, National High School of Arts and Crafts, Electrical Engineering Department, E2SN Research Laboratory, Rabat, Morocco

Abstract. *Transfer Learning (TL) is a popular deep learning technique used in medical image analysis, especially when data is limited. It leverages pre-trained knowledge from State-Of-The-Art (SOTA) models and applies it to specific applications through Fine-Tuning (FT). However, fine-tuning large models can be time-consuming, and determining which layers to use can be challenging. This study explores different fine-tuning strategies for five SOTA models (VGG16, VGG19, ResNet50, ResNet101, and InceptionV3) pre-trained on ImageNet. It also investigates the impact of the classifier by using a linear SVM for classification. The experiments are performed on four open-access ultrasound datasets related to breast cancer, thyroid nodules cancer, and salivary glands cancer. Results are evaluated using a five-fold stratified cross-validation technique, and metrics like accuracy, precision, and recall are computed. The findings show that fine-tuning 15% of the last layers in ResNet50 and InceptionV3 achieves good results. Using SVM for classification further improves overall performance by 6% for the two best-performing models. This research provides insights into fine-tuning strategies and the importance of the classifier in transfer learning for ultrasound image classification.*

Keywords: CNN, transfer learning, fine-tuning, SVM, ultrasound images, cancer classification

OPTYMALIZACJA KLASYFIKACJI OBRAZÓW ULTRASONOGRAFICZNYCH TECHNIKĄ TRANSFER LEARNING: STRATEGIE DOSTRAJANIA I WPŁYW KLASYFIKATORA NA WSTĘPNIE WYTRENOWANE WARSTWY WEWNĘTRZNE

Streszczenie. *Transfer Learning (TL) to popularna technika głębokiego uczenia stosowana w analizie obrazów medycznych, zwłaszcza gdy ilość danych jest ograniczona. Wykorzystuje ona wstępnie wyszkoloną wiedzę z modeli State-Of-The-Art (SOTA) i zastosowanie ich do konkretnych aplikacji poprzez dostrajanie (Fine-Tuning – FT). Jednak dostrajanie dużych modeli może być czasochłonne, a określenie, których warstw użyć, może stanowić wyzwanie. W niniejszym badaniu przeanalizowano różne strategie dostrajania dla pięciu modeli SOTA (VGG16, VGG19, ResNet50, ResNet101 i InceptionV3) wstępnie wytrenowanych na ImageNet. Zbadano również wpływ klasyfikatora przy użyciu liniowej SVM do klasyfikacji. Eksperymenty przeprowadzono na czterech ogólnodostępnych zbiorach danych ultrasonograficznych związanych z rakiem piersi, rakiem guzków tarczycy i rakiem gruczołów ślinowych. Wyniki są oceniane przy użyciu techniki pięciowarstwowej walidacji krzyżowej, a wskaźniki takie jak dokładność, precyzja i odzyskiwanie są obliczane. Wyniki pokazują, że dostrojenie 15% ostatnich warstw w ResNet50 i InceptionV3 osiąga dobre wyniki. Użycie SVM do klasyfikacji dodatkowo poprawia ogólną wydajność o 6% dla dwóch najlepszych modeli. Badania te zapewniają informacje na temat strategii dostrajania i znaczenia klasyfikatora w uczeniu transferowym dla klasyfikacji obrazów ultrasonograficznych.*

Słowa kluczowe: CNN, transfer learning, dostrajanie, SVM, obrazy ultrasonograficzne, klasyfikacja nowotworów

Introduction

Medical ultrasound imaging is a widely used modality for diagnosing various conditions, such as tumors, cysts, and abnormalities in organs and tissues. It is a non-invasive technique, less expensive, and can provide real-time images for diagnosis purposes [15].

Accurate classification of ultrasound medical images plays a crucial role in the clinical decision-making process. However, it can be challenging due to the complexity of these images, as well as the limited availability of annotated data for training Deep Learning (DL) models based on Convolutional Neural Networks (CNNs) [7, 21]. These kinds of models have shown remarkable success in image classification tasks, but they may face limitations in medical ultrasound images due to the previous-mentioned obstacles.

In recent years, Transfer Learning (TL) has shown promising results in various computer vision tasks, and become a prominent DL technique allowing models trained on large datasets such as ImageNet, to be fine-tuned on smaller target datasets. TL has the potential to address the data deficiency limitations found in medical images, making it a valuable tool for improving the accuracy and efficiency of Computer-Aided-Diagnosis (CAD) systems.

TL pre-trained models have already learned low-level and generic features common to many images such as edges, contours, shapes, and so on. While high-level features specific to the classification task are learned through the classifier by means of Fine-Tuning (FT). This technique can help reduce the computational cost of DL models, saving the effort of building layers from scratch. This can be especially beneficial in the medical field where annotated data are scarce and costly to obtain.

Despite its advantages, TL requires essentially a classifier built on-top for FT purposes. The classifier can be as simple as a Global Max-Pooling (GMP) or Flatten layer with Fully Connected (FC) layers that match the number of target classes. Utilizing such a shallow Multilayer Perceptron (MLP) classifier may not guarantee optimal results and can, in certain cases, result in overfitting the training set due to the depth of the pre-trained models. However, choosing an adequate FT strategy and a questionable classifier can help overcome these limitations. In such cases, hybrid approaches combining CNNs extracted features along with Machine Learning (ML) estimators can offer improved performance and robustness.

In recent years, there has been growing interest in combining CNNs with other classifiers to improve accuracy, robustness, and interpretability. One popular hybrid approach is the combination of CNN and Support Vector Machines (SVM) [4].

CNNs are known for their ability to learn hierarchical features from images, automatically capturing relevant patterns and structures. SVM, on the other hand, is a well-known ML estimator that was initially designed for binary classification tasks providing a clear decision boundary while handling small datasets effectively. Combining the feature extraction capabilities of CNNs with the discriminative power of SVM, this approach can lead to a better-generalized classification outcome.

In this paper, we emphasize the significance of FT techniques in TL for the classification of ultrasound images. We evaluate five of the classically recognized ImageNet pre-trained models, namely VGG16, VGG19, ResNet50, ResNet101, and InceptionV3, using various FT strategies. Additionally, we investigate the effectiveness of the hybrid CNN-SVM approach by incorporating an SVM classifier on top of the fine-tuned pre-trained models. The findings of this study provide valuable insights into the advantages of fine-tuning specific layers and the hybrid CNN-SVM approach for ultrasound image classification.

1. Related works

In recent research, TL has been widely used to address the demand for large labeled data required to train DL models. Training these models from scratch for medical images can be challenging due to several reasons. Notably, dataset availability is often expensive in collection and storage, particularly when involving professional radiologists in the annotation process which is time-consuming and error-prone posing challenges in training DL models. As a result, researchers have introduced TL as an efficient and low-cost technique to remediate the lack of data. Many researchers have used ImageNet pre-trained models, such as VGG16, ResNet, and InceptionV3 in various medical imaging applications, such as skin cancer, breast cancer, and so on.

TL commonly employs two strategies: fine-tuning, utilized when the target dataset is enough for the training, and feature extraction, an alternative method for leveraging low-level features from pre-trained models. Besides, data augmentation techniques such as rotation, cropping, noise adding, and color manipulation

have also been commonly used to expand the dataset and prevent overfitting [8, 21].

TL has made major contributions to medical image analysis by overcoming the problem of data scarcity and saving time and hardware resources. In the review paper [5], authors investigated 121 studies around selecting the best backbone models and TL approaches for medical images. They have therefore divided TL strategies into four categories including feature extractor, feature extractor hybrid, fine-tuning, and fine-tuning from scratch. Authors have also declared Inception, ResNet, VGG, AlexNet, and LeNet as the most common TL models used in literature. The same was also confirmed in the review paper [6]. Additionally, authors of [5] have recommended preferring ResNet and Inception models as feature extractors due to their performance and computational efficiency.

In the same context, authors in [14] suggest a novel deep learning cascaded feature framework to address the issue of the high dimensionality of features extracted from deep layers of pre-trained CNNs models. Their framework utilized pre-trained models such as AlexNet, VGG, and GoogleNet to extract shallow and deep features, and have employed their univariate strategy to overcome the dimensionality and multicollinearity issues in the extracted features. The evaluation of their proposed framework yielded an accuracy of 98.50%, sensitivity of 98.06%, specificity of 98.99%, and precision of 98.98%.

The authors of the paper [3], investigated ten TL models as backbones for the U-Net [12] model for segmenting breast ultrasound images. The obtained results demonstrated the efficiency of pre-trained models in extracting relevant features for breast lesions segmentation.

In [19], authors used deep learning pre-trained models such as ResNet50, DenseNet121, and EfficientNetB3, besides transformer-based methods such as ViT-B/16 to classify Salivary gland tumors. These tumors are commonly inferred from the parotid glands where only 20% of the tumors are malignant. Authors have performed a binary classification on a dataset of 251 patients, with about 29.5% of the malignant cases. They used data augmentation techniques during the training such as random flipping, rotating, blurring, and lighting adjustments. Their results outperformed those of inexperienced radiologists. Notably, EfficientNetB3 and DenseNet121 models achieved accuracy and Area Under the Curve (AUC) of 80%, 0.82, and 77%, 0.81 respectively. In the same context, [20] have trained a modified ResNet18 network over 1200 epochs with a learning rate of 1e-6 using Adam optimizer on a dataset of parotid lesions from 232 patients and a total of 3791 cropped parotid gland region images. The dataset was partitioned into 90% of training and 10% of validation sets. Additionally, data augmentation techniques such as image flipping and contrast adjustment were used for data enhancement. Authors have reported an accuracy of 82.18% with a micro-AUC of 0.93.

In the context of hybrid models, authors in [1] and [17] demonstrated the advantages of using linear L2-SVM as a top layer instead of softmax in DL architectures, highlighting benefits such as differentiability and stronger error penalization. They showed that L2-SVM is slightly better than L1-SVM, and they used linear SVMs in their experiments. They tested their approach on well-known datasets and achieved competitive results in a facial expression recognition competition. They highlighted the effectiveness of the last layer SVM in comparison to softmax, and they attributed the performance gain to the superior regularization effects of the SVM loss function rather than better parameter optimization.

The work in [9] develops a generic CAD system based on features extracted from pre-trained CNNs tested on 12 open-access image datasets. The authors aimed to explore the power of intermediate and last layers of ImageNet pre-trained models such as GoogleNet (Inception), ResNet, and DenseNet201 feature vectors for training an ensemble of SVMs. The extracted features are fed to SVMs, then combined for the final results.

Regarding thyroid nodule classification, [16] suggested a hybrid model based on CNN and SVM, compiled with a hinge loss function. They evaluated the results on two public datasets containing 1180 and 2616 thyroid ultrasound images after applying data augmentation. They reported an accuracy of 94.57%, 96% and specificity of 91.89%, 93.93%, and a sensitivity of 96.70%, 97.80% for both datasets, dataset-1 and dataset-2 respectively.

To guarantee optimal performances, SVM hyperparameters such as the C penalty parameter, and the kernel were investigated. [18] used Quantum-Behaved Particle Swarm Optimization (QPSO) algorithm to optimize SVM parameters due to its global search ability and fewer control parameters. A hybrid model consisting of a LeNet-5 network and SVM was utilized and validated on breast cancer cell images. The model achieved a test accuracy of 93.15%, outperforming the model without SVM, by 1.9%.

2. Proposed method

In this section, we introduce and discuss the different phases involved in this study, with a particular emphasis on FT strategies. Then, we delve into the details of the FT hybrid CNN-SVM approach and explore its potential in enhancing the overall classification performances. The study workflow is described in Fig. 1.

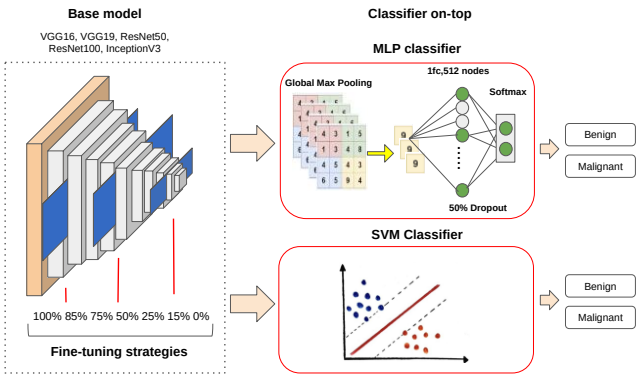


Fig. 1. Study workflow

2.1. Study workflow

To evaluate the experiment, four publicly available ultrasound datasets were used representing diverse anatomical regions and clinical cancer scenarios. Breast, thyroid nodules, and salivary glands are the organs investigated with their corresponding datasets 1) Breast Mendeley [11], 2) Breast Ultrasound Images (BUSI) [2], 3) Digital Database of Thyroid Images (DDTI) [10], and 4) Salivary Glands from Ultrasound cases website [13].

TL FT strategies were separately applied to each dataset using four state-of-the-art pre-trained models: VGG16, VGG19, ResNet50, ResNet101, and InceptionV3.

A model selection mechanism was implemented using a 5-fold Stratified Cross-Validation (SCV) technique, where the best model was saved during each fold for subsequent evaluation. This technique is widely recognized and effective in handling imbalanced datasets, which is often the case for medical images. Additionally, classification evaluation metrics including accuracy, precision, recall, and AUC values were computed. Furthermore, the Receiver Operator Characteristic (RoC) curve and Confusion Matrix (CM) were provided for a comprehensive description of the classification outcomes for the two best-performing models.

FT strategies and the effectiveness of the hybrid CNN-SVM approach were separately evaluated for each dataset. This enabled a thorough evaluation of the proposed approaches within the specific medical ultrasound domain, offering valuable insights into their performance and applicability. Figure 2 displays four samples from each dataset, while table 1 provides an overview of their class distribution.

2.2. Models configurations

A. Phase 1: Transfer learning fine-tuning approach (FT)

This approach aims at extracting optimal features from ultrasound images. For this task, pre-trained models with ImageNet weights were used as backbones, followed by a shallow MLP classifier with a GlobalMaxPooling layer, a fully connected layer (512 nodes) with 50% dropout, and a final fully connected layer (2 nodes) with softmax activation function.

Table 2 describes the seven fine-tuning strategies that were implemented, irrespective of the models' depth, and highlights the number of total and frozen layers for the investigated pre-trained models.

Models in this phase were compiled using binary cross-entropy and Adam optimizer, then trained for 30 epochs with a learning rate of 0.0001 and a batch size of 4. Models evaluation was conducted using a Stratified Cross-Validation technique with metrics computed and averaged across the 5-folds representing the overall model performance.

B. Phase 2: Transfer learning fine-tuning hybrid approach (FT-SVM)

This approach consisted of employing SVM in replacement of the basic classifier made earlier. SVM inputs were obtained from the best features extracted from the FT models that were saved during the earlier phase.

Based on the literature review, the linear kernel-based SVM classifier has proved its efficiency in delivering good results in image classification tasks. Therefore, only SVM's C-hyperparameter was tuned using a GridSearch mechanism with a 5-fold SCV technique for values ranging from 0.1, 1, 10, and 100. SVM's C parameter is known to penalize each misclassified point and controls the SVM decision boundaries. Thus, choosing the right C value may require more thorough tests, hence the used GridSearch mechanism.

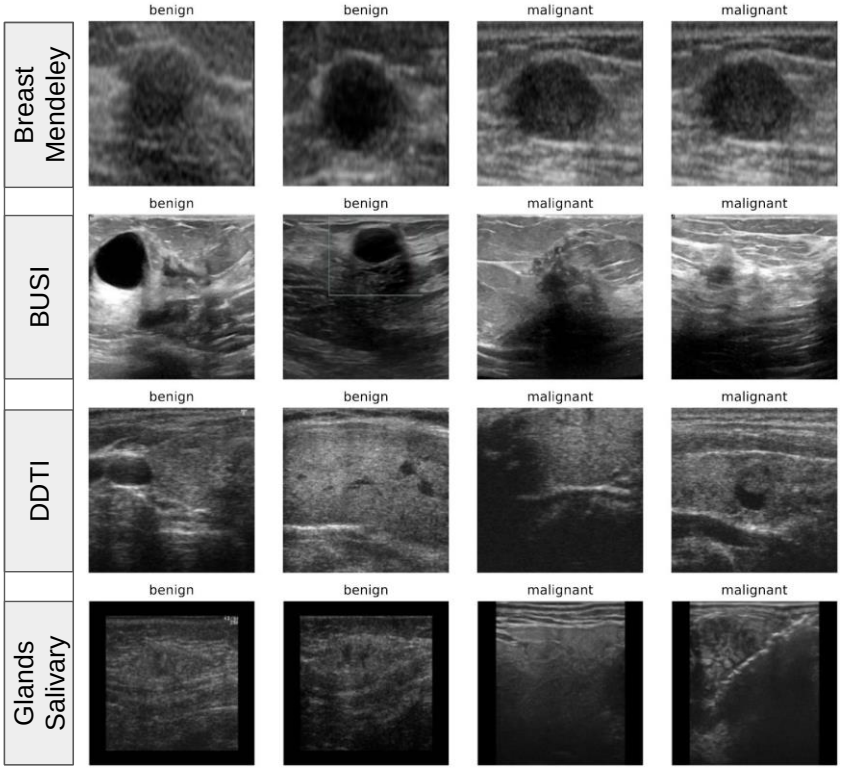


Fig. 2. Ultrasound datasets overview

Table 1. Datasets class distribution

Dataset	Benign	Malignant	Total
Breast Mendeley	100	150	250
Breast BUSI	437	210	647
Thyroid DDTI	49	399	448
Salivary Glands	781	109	890

Table 2. Fine-tuning strategies and pre-trained models: layers description

	VGG16	VGG19	ResNet50	ResNet101	InceptionV3
Total layers	19	22	175	345	311
FT strategies (%)	Number of frozen layers				
0% (purely TL)	19	22	175	345	311
15%	16	19	149	293	264
25%	14	16	131	259	233
50%	10	11	88	173	156
75%	5	6	44	86	78
85%	3	4	26	52	47
100% (scratch fine-tuning)	0	0	0	0	0

3. Experimental results and discussion

The investigated fine-tuning strategies yielded the results described in table 3 Since there are many metrics involved, we have structured the results in this table by the AUC value which reflects the true positive and the false negative rates of the classification. The AUC values were computed for each strategy, both for the fine-tuned models using a basic MLP classifier and the fine-tuned models using SVM. The best-performing models were chosen based on their AUC values performances across all models. Fine-tuning 15% of layers was found the best strategy and resulted in a good classification, specifically when performed by SVM. Table 4 shows the computed metrics for the five models along with the optimal SVM-C parameter value. Results in this table are structured as follows: each metric consists of a pair of values, representing the performance of the fine-tuned layers with the basic classifier and the fine-tuned layers with SVM.

For each dataset, excluding Breast Mendeley due to its potentially optimistic results that could impact the overall accuracy of the study, the confusion matrix and RoC curves were generated separately for the winning strategy in both ResNet50 and InceptionV3. Additionally, in the breast BUSI dataset, ResNet50 showed an increase of 18% in sensitivity, while InceptionV3 exhibited an increase of 12%. For the thyroid DDTI dataset, ResNet50 achieved a sensitivity increase of 3%, whereas InceptionV3 showed a 2% increase. In the case of the salivary glands dataset, ResNet50 demonstrated an increase of 0.52% in sensitivity compared to InceptionV3 with a 0.42%. Besides, InceptionV3 was found the most consistent model across all strategies and different ultrasound datasets being used. Figures 3, 5, and 7 display the confusion matrix of ResNet50 and InceptionV3 for the breast BUSI, thyroid DDTI, and Salivary Glands datasets, respectively, using the 15% FT strategy. Additionally, Figures 4, 6 and 8 illustrate the differences in RoC curves and emphasize the superior performance of SVM in the overall classification.

Table 3. Fine-tuning strategies by AUC values: FT and FT-SVM

AUC (%)	VGG16	VGG19	ResNet50	ResNet101	InceptionV3
Breast Mendeley	0% :(0.97, 0.95) 15% :(0.94, 1.00) 25% :(0.86, 1.00) 50% :(1.00, 1.00) 75% :(1.00, 1.00) 85% :(1.00, 0.98) 100%:(1.00, 1.00)	0% :(0.94, 0.97) 15% :(0.76, 0.90) 25% :(0.90, 1.00) 50% :(0.94, 1.00) 75% :(1.00, 0.98) 85% :(1.00, 0.99) 100%:(0.97, 1.00)	0% :(0.53, 0.70) 15% :(1.00, 1.00) 25% :(1.00, 1.00) 50% :(0.99, 1.00) 75% :(0.66, 0.97) 85% :(0.66, 0.99) 100%:(1.00, 1.00)	0% :(0.74, 0.97) 15% :(1.00, 1.00) 25% :(1.00, 1.00) 50% :(0.95, 0.95) 75% :(0.95, 1.00) 85% :(0.97, 1.00) 100%:(1.00, 1.00)	0% :(0.99, 1.00) 15% :(1.00, 1.00) 25% :(1.00, 1.00) 50% :(1.00, 1.00) 75% :(1.00, 1.00) 85% :(1.00, 0.99) 100%:(1.00, 1.00)
Breast BUSI	0% :(0.80, 0.81) 15% :(0.76, 0.93) 25% :(0.80, 0.90) 50% :(0.84, 0.94) 75% :(0.86, 0.96) 85% :(0.87, 0.95) 100%:(0.64, 0.88)	0% :(0.78, 0.84) 15% :(0.67, 0.85) 25% :(0.79, 0.88) 50% :(0.78, 0.96) 75% :(0.84, 0.95) 85% :(0.86, 0.94) 100%:(0.60, 0.87)	0% :(0.51, 0.78) 15% :(0.87, 0.98) 25% :(0.85, 0.93) 50% :(0.81, 0.95) 75% :(0.76, 0.89) 85% :(0.80, 0.90) 100%:(0.90, 0.99)	0% :(0.56, 0.76) 15% :(0.81, 0.95) 25% :(0.82, 0.95) 50% :(0.75, 0.81) 75% :(0.76, 0.83) 85% :(0.78, 0.85) 100%:(0.84, 0.97)	0% :(0.81, 0.84) 15% :(0.90, 0.99) 25% :(0.90, 0.98) 50% :(0.88, 0.98) 75% :(0.83, 0.97) 85% :(0.85, 0.96) 100%:(0.88, 0.98)
Thyroid DDTI	0% :(0.51, 0.50) 15% :(0.50, 0.50) 25% :(0.50, 0.50) 50% :(0.52, 0.50) 75% :(0.50, 0.50) 85% :(0.50, 0.55) 100%:(0.50, 0.50)	0% :(0.50, 0.50) 15% :(0.50, 0.50) 25% :(0.50, 0.50) 50% :(0.50, 0.50) 75% :(0.50, 0.50) 85% :(0.51, 0.59) 100%:(0.50, 0.50)	0% :(0.50, 0.50) 15% :(0.59, 0.94) 25% :(0.52, 0.83) 50% :(0.50, 0.50) 75% :(0.50, 0.50) 85% :(0.50, 0.51) 100%:(0.55, 0.50)	0% :(0.50, 0.50) 15% :(0.49, 0.50) 25% :(0.50, 0.50) 50% :(0.50, 0.50) 75% :(0.50, 0.50) 85% :(0.50, 0.50) 100%:(0.54, 0.57)	0% :(0.52, 0.50) 15% :(0.54, 0.91) 25% :(0.55, 0.92) 50% :(0.55, 0.50) 75% :(0.51, 0.66) 85% :(0.53, 0.62) 100%:(0.51, 0.93)
Salivary Glands	0% :(0.61, 0.50) 15% :(0.50, 0.50) 25% :(0.56, 0.75) 50% :(0.50, 0.50) 75% :(0.72, 0.84) 85% :(0.68, 0.88) 100%:(0.50, 0.50)	0% :(0.54, 0.50) 15% :(0.50, 0.50) 25% :(0.50, 0.50) 50% :(0.55, 0.50) 75% :(0.58, 0.89) 85% :(0.65, 0.62) 100%:(0.50, 0.50)	0% :(0.50, 0.50) 15% :(0.67, 0.94) 25% :(0.60, 0.61) 50% :(0.50, 0.55) 75% :(0.50, 0.50) 85% :(0.50, 0.50) 100%:(0.79, 0.96)	0% :(0.66, 0.50) 15% :(0.61, 0.93) 25% :(0.52, 0.50) 50% :(0.50, 0.50) 75% :(0.50, 0.50) 85% :(0.50, 0.66) 100%:(0.79, 0.95)	0% :(0.58, 0.50) 15% :(0.71, 0.95) 25% :(0.74, 0.90) 50% :(0.80, 0.96) 75% :(0.70, 0.97) 85% :(0.63, 0.85) 100%:(0.79, 0.94)

Table 4. Fine-tuning results for the 15% layers

Dataset	Metrics(%)	VGG16	VGG19	ResNet50	ResNet101	InceptionV3
Breast Mendeley	SVM-C parameter	1	0.1	0.1	0.1	1
	Accuracy	(0.9480, 1.0000)	0.7960, 0.9160)	(1.0000, 1.0000)	(1.0000, 1.0000)	(1.0000, 1.0000)
	Precision	(0.9419, 1.0000)	(0.7619, 0.8778)	(1.0000, 1.0000)	(1.0000, 1.0000)	(1.0000, 1.0000)
	Sensitivity	(0.9733, 1.0000)	(0.9600, 1.0000)	(1.0000, 1.0000)	(1.0000, 1.0000)	(1.0000, 1.0000)
Breast BUSI	SVM-C parameter	100	100	1	0.1	10
	Accuracy	(0.8176, 0.9412)	(0.7650, 0.8654)	(0.8887, 0.9768)	(0.8469, 0.9644)	(0.9072, 0.9876)
	Precision	(0.7875, 0.9234)	(0.7735, 0.7996)	(0.8415, 0.9494)	(0.8100, 0.9673)	(0.8504, 0.9764)
	Sensitivity	(0.6000, 0.8952)	(0.3904, 0.7904)	(0.8095, 0.9809)	(0.6904, 0.9238)	(0.8666, 0.9857)
Thyroid DDTI	SVM-C parameter	0.1	0.1	10	0.1	100
	Accuracy	(0.8893, 0.8906)	(0.8893, 0.8906)	(0.8849, 0.9821)	(0.8783, 0.8906)	(0.8783, 0.9753)
	Precision	(0.8893, 0.8906)	(0.8893, 0.8906	(0.9088, 0.9852)	(0.8881, 0.8906)	(0.8970, 0.9778)
	Sensitivity	1.0000, 1.0000)	(1.0000, 1.0000)	(0.9676, 0.9949)	(0.9875, 1.0000)	(0.9751, 0.9949)
Salivary Glands	SVM-C parameter	0.1	0.1	10	1	10
	Accuracy	(0.8786, 0.8775)	(0.8775, 0.8775)	(0.8932, 0.9775)	(0.8853, 0.9662)	(0.8876, 0.9752)
	Precision	(1.0000, 0.0000)	(0.0000, 0.0000)	(0.6060, 0.9345)	(0.5744, 0.8545)	(0.5463, 0.8904)
	Sensitivity	(0.0091, 0.0000)	(0.0000, 0.0000)	(0.3669, 0.8809)	(0.2477, 0.8722)	(0.4862, 0.9090)

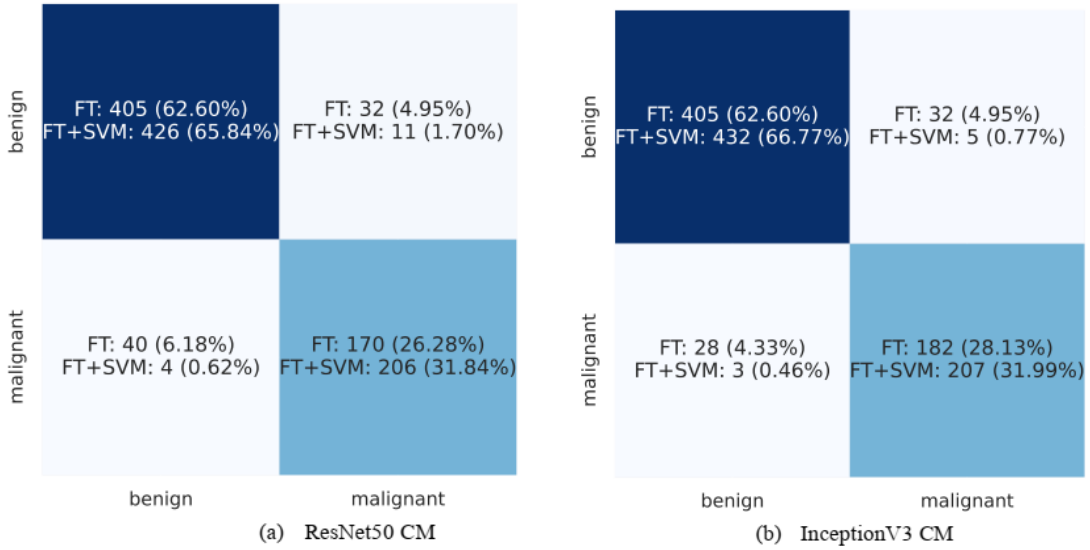


Fig. 3. BUSI confusion matrix curves: 15% of layers FT and FT+SVM

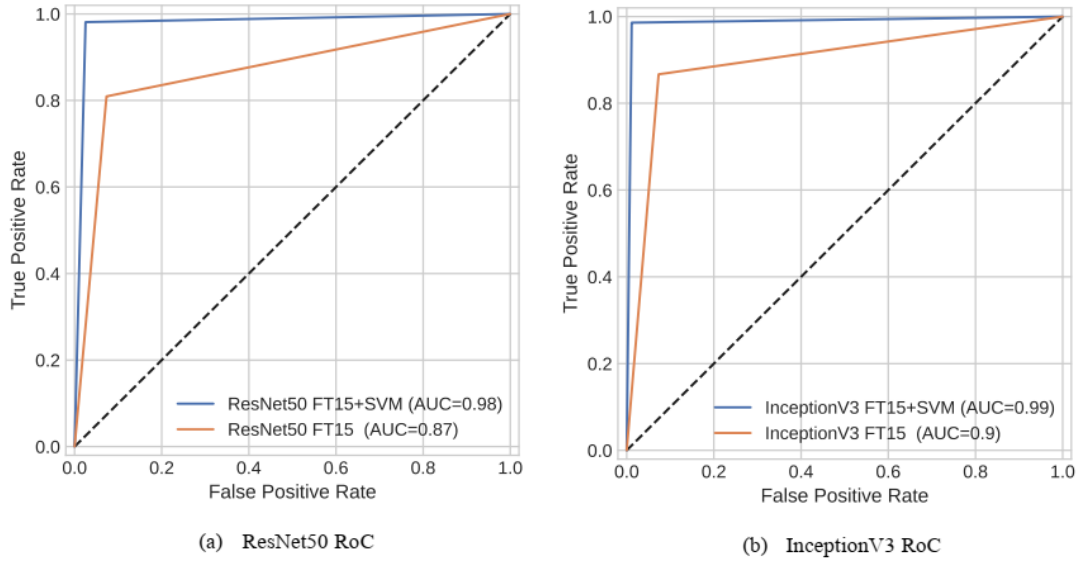


Fig. 4. BUSI RoC curve: 15% of layers FT and FT+SVM

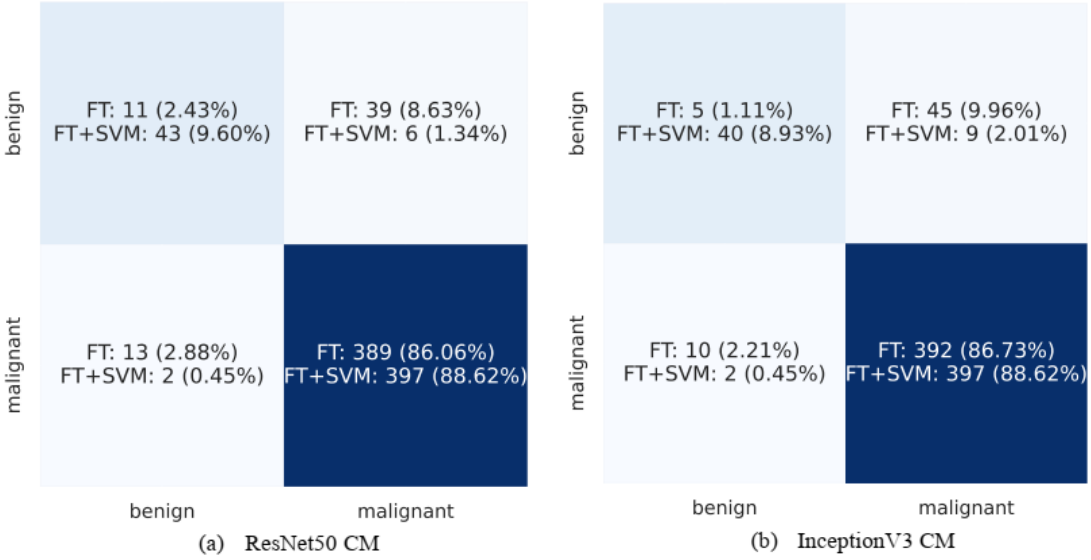


Fig. 5. DDTI confusion matrix: 15% of layers FT and FT+SVM

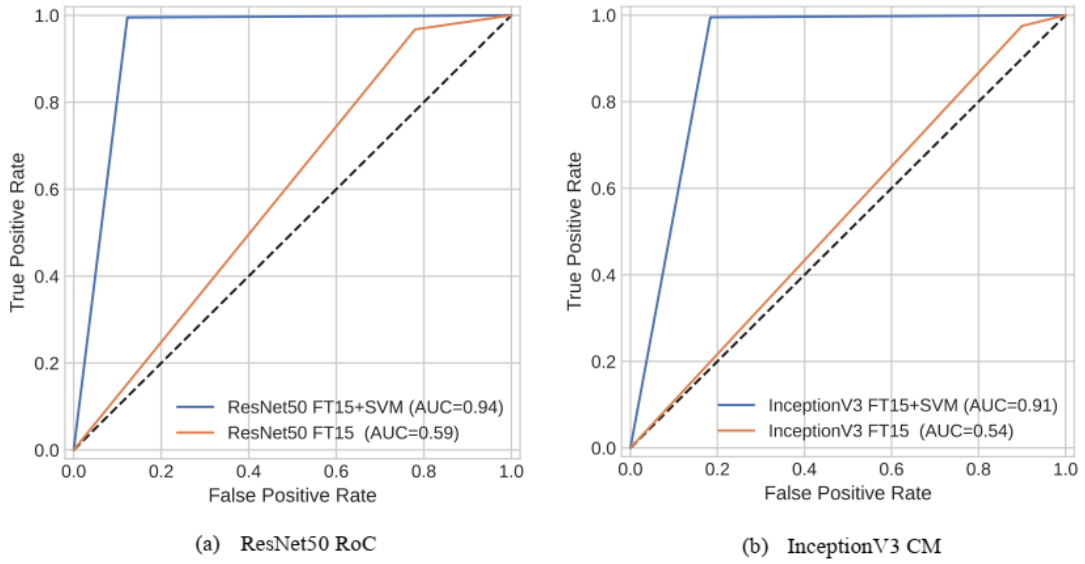


Fig. 6. DDTI RoC curves: 15% of layers FT and FT+SVM

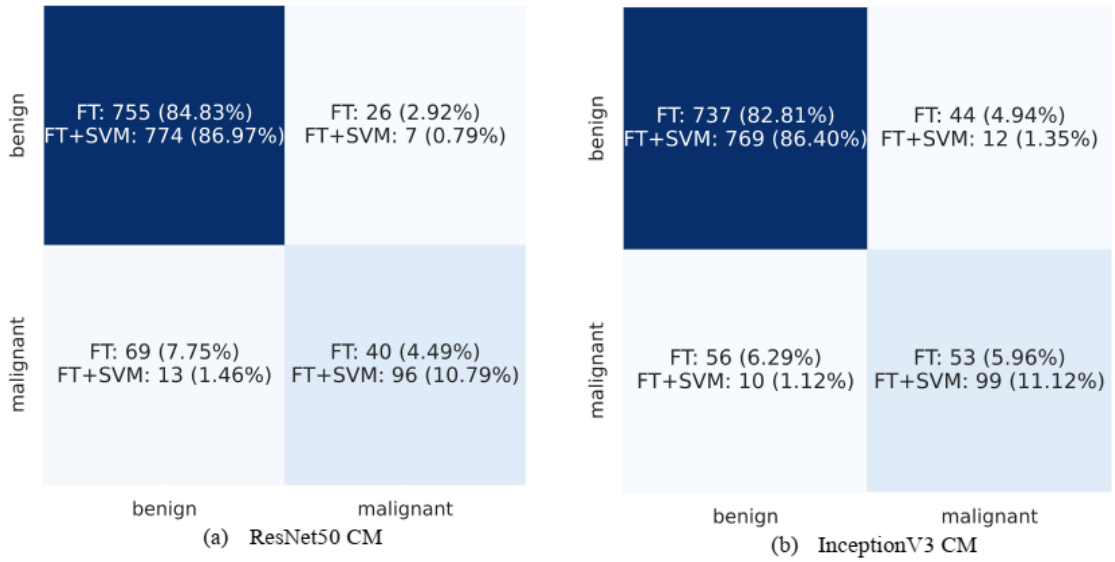


Fig. 7. Salivary Glands confusion matrix: 15% of layers FT and FT+SVM

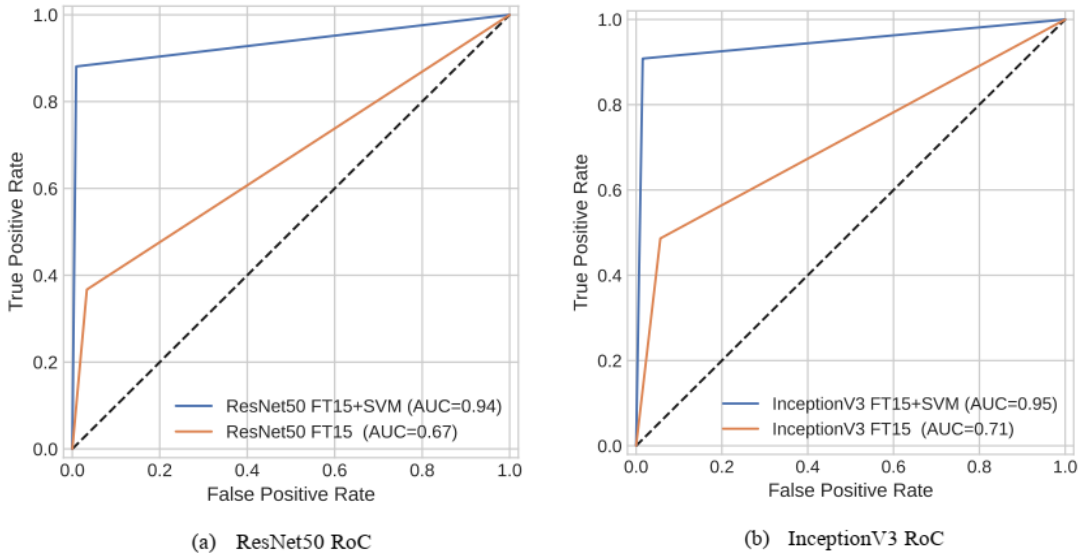


Fig. 8. Salivary Glands RoC curves: 15% of layers FT and FT+SVM

4. Conclusion

This paper demonstrates the effectiveness of transfer learning and the effect of fine-tuning specific layers in enhancing ultrasound image classification tasks. Using different state-of-the-art ImageNet pre-trained models, various fine-tuning strategies were implemented and investigated such as fine-tuning 0%, 15%, 25%, 50%, 75%, 85%, and 100% of inner layers. The evaluation of these strategies was performed in two phases, first, fine-tuning with a basic classifier, and second, replacing the classifier with a linear SVM. The whole evaluation process has been implemented with a 5-fold cross-validation mechanism ensuring robust model evaluation and selection. Among the five implemented models, two models namely, ResNet50, and InceptionV3 have shown good performances while fine-tuning 15% of their layers. Additionally, the overall performance of these models has increased significantly while adopting a hybrid approach by leveraging a linear SVM on the classifier part of the fine-tuned models. The results of this study underscore the importance of optimizing deep learning techniques in ultrasound image analysis. Additionally, they shed light on the significance of fine-tuning strategies and classifier selection in achieving accurate and reliable classification outcomes.

Conflict of interest

No author claims to have any conflicts of interest.

References

[1] Agarap A. F.: An architecture combining convolutional neural network (cnn) and support vector machine (svm) for image classification. arXiv preprint: 1712.03541, 2017.
[2] Al-Dhabyani W. et al.: Dataset of breast ultrasound images. Data Brief. 28, 2019, 104863.
[3] Bal-Ghaoui M. et al.: U-net transfer learning backbones for lesions segmentation in breast ultrasound images. International Journal of Electrical and Computer Engineering (IJECE) 13, 2023, 5747, [http://doi.org/10.11591/ijece.v13i5.pp5747-5754].
[4] Cortes C., Vapnik V.: Support-vector networks. Machine learning 20, 1995, 273–297.
[5] Kim H. E. et al.: Transfer learning for medical image classification: A literature review. BMC medical imaging 22(1), 2022, 69.
[6] Kora P. et al.: Transfer learning techniques for medical image analysis: A review. Biocybernetics and Biomedical Engineering 42(1), 2022, 79–107.
[7] LeCun Y. et al.: Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems 2, 1989.
[8] Mukhlif A. A. et al.: An extensive review of state-of-the-art transfer learning techniques used in medical imaging: Open issues and challenges. Journal of Intelligent Systems 31(1), 2022, 1085–1111.
[9] Nanni L., Ghidoni S., Brahnam S.: Deep features for training support vector machines. Journal of Imaging 7(9), 2021, 177.
[10] Pedraza L. et al.: An open access thyroid ultrasound image database. 10th International symposium on medical information processing and analysis 9287, 2015, 188–193.
[11] Rodrigues P. S.: Breast ultrasound image. Mendeley Data 1(10), 2017, 17632.
[12] Ronneberger O., Fischer P., Brox T.: U-net: Convolutional networks for biomedical image segmentation. 18 International Conference Medical Image Computing and Computer-Assisted Intervention – MICCAI, Munich, 2015, 234–241.
[13] Salivary Glands Ultrasound Cases. Website [https://www.ultrasoundcases.info/cases/head-and-neck/salivary-glands/] (accessed: April 15, 2023).
[14] Samee N. A. et al.: Deep learning cascaded feature selection framework for breast cancer classification: Hybrid cnn with univariate-based approach. Mathematics 10(19), 2022, 3631.
[15] Shung K. K.: Diagnostic ultrasound: Past, present, and future. Journal of Medical and Biological Engineering 31(6), 2011, 371–374.
[16] Srivastava R., Kumar P.: A cnn-svm hybrid model for the classification of thyroid nodules in medical ultrasound images. International Journal of Grid and Utility Computing 13(6), 2022, 624–639.
[17] Tang Y.: Deep learning using linear support vector machines. arXiv, preprint: 1306.0239, 2013.
[18] Wang Y. et al.: A hybrid classification method of medical image based on deep learning. Research Square, preprint, 2021.
[19] Wang Y. et al.: The diagnostic value of ultrasound-based deep learning in differentiating parotid gland tumors. Journal of Oncology, 2022.
[20] Xia X. et al.: Deep learning for differentiating benign from malignant parotid lesions on mri images. Frontiers in Oncology 11, 2021, 632104.
[21] Yu X. et al.: Transfer learning for medical images analyses: A survey. Neurocomputing 489, 2022, 230–254.

M.Sc. Eng. Mohamed Bal-Ghaoui
e-mail: mohamed_balghaoui@um5.ac.ma

Received his Engineer's degree in Biomedical Engineering at the University of Mohammed V, ENSAM Rabat. He is a Ph.D. candidate researcher in Artificial Intelligence applied to Biomedical engineering. His research interest includes AI applied to Ultrasound Images for the early detection of cancer. He is part of E2SN Team – Biomedical Engineering Research Laboratory at ENSAM-Rabat.

http://orcid.org/0000-0002-2143-6458

Prof. My Hachem El Yousfi Alaoui
e-mail: h.elyousfi@um5r.ac.ma

Professor of Biomedical Engineering at the University of Mohammed V, ENSAM-Rabat, he is a member of the research laboratory E2SN – Biomedical Engineering Research Laboratory at ENSAM-Rabat, Mohamed V University in Rabat, Morocco. Members of the E2SN research group. Prof. El Yousfi is current research work is focused on biomedical data processing, AI, IoT and the hardware implementation of associated circuits.

http://orcid.org/0000-0003-4285-0540

Prof. Abdelilah Jilbab
e-mail: a.jilbab@um5r.ac.ma

Professor of Electrical Engineering at ENSAM Rabat of the University of Mohammed V in Rabat. He acquired his Ph.D. in Computer and Telecommunication from Mohammed V Agdal University, Rabat, Morocco in February 2009. He has published in the field of image processing, sensor networks, and signal processing for Parkinson's disease. His current interest is in embedded systems and wireless sensor networks (WSN) applied to biomedical. Dr. Jilbab is a member of the research laboratory E2SN – Biomedical Engineering Research Laboratory at ENSAM-Rabat at Mohamed V University.

http://orcid.org/0000-0002-1577-9040

Prof. Abdennaser Bourouhou
e-mail: a.bourouhou@um5r.ac.ma

Professor of Electrical Engineering at ENSAM, Mohammed V University Rabat. He acquired his Ph.D. in Electronics and Signal Processing from Ibn Tofail University in Kenitra, Morocco in April 2008. A. Bourouhou is a member of the Electronic Systems Sensors and Nanobiotechnology (E2SN) team of the STIS research center of UM5 Rabat. His research interests include biosignal processing, wireless sensor networks (WSNs) for environmental protection, and medical monitoring.

http://orcid.org/0000-0002-6150-5374

http://doi.org/10.35784/iapgos.3783 *received: 23.06.2023 | revised: 09.08.2023 | accepted: 27.11.2023 | available online: 20.12.2023*

A GENERATIVE MODEL FOR DEEP FAKE AUGMENTATION OF PHONOCARDIOGRAM AND ELECTROCARDIOGRAM SIGNALS USING LSGAN AND CYCLE GAN

Swarajya Madhuri Rayavarapu[1], Tammineni Shanmukha Prasanthi[1], Gottapu Santosh Kumar[2], Gottapu Sasibhushana Rao[1], Gottapu Prashanti[3]

[1]Andhra University, Department of Electronics and Communication Engineering, Visakhapatnam, India, [2]Gayatri Vidya Parishad College of Engineering, Department of Civil Engineering, Visakhapatnam, India, [3]Avanthi Institute of Pharmaceutical Sciences, Department of Pharmaceutical Technology, Vizianagaram, India

Abstract*. In order to diagnose a range of cardiac conditions, it is important to conduct an accurate evaluation of either phonocardiogram (PCG) and electrocardiogram (ECG) data. Artificial intelligence and machine learning-based computer-assisted diagnostics are becoming increasingly commonplace in modern medicine, assisting clinicians in making life-or-death decisions. The requirement for an enormous amount of information for training to establish the framework for a deep learning-based technique is an empirical challenge in the field of medicine. This increases the risk of personal information being misused. As a direct result of this issue, there has been an explosion in the study of methods for creating synthetic patient data. Researchers have attempted to generate synthetic ECG or PCG readings. To balance the dataset, ECG data were first created on the MIT-BIH arrhythmia database using LS GAN and Cycle GAN. Next, using VGGNet, studies were conducted to classify arrhythmias for the synthesized ECG signals. The synthesized signals performed well and resembled the original signal and the obtained precision of 91.20%, recall of 89.52% and an F1 score of 90.35%.*

Keywords: arrhythmia, auscultation, electrocardiogram, phonocardiogram, generative networks

GENERATYWNY MODEL Z DEEP FAKE AUGUMENTATION DLA SYGNAŁÓW Z FONOKARDIOGRAMU ORAZ ELEKTROKARDIOGRAMU W STRUKTURACH LSGAN ORAZ CYCLE GAN

Streszczenie. *W celu zdiagnozowania szeregu chorób serca, istotne jest przeprowadzenie dokładnej oceny danych z fonokardiogramu (PCG) i elektrokardiogram (EKG). Sztuczna inteligencja i diagnostyka wspomagana komputerowo, oparta na uczeniu maszynowym stają się coraz bardziej powszechne we współczesnej medycynie, pomagając klinicystom w podejmowaniu krytycznych decyzji. Z kolei, Wymóg ogromnej ilości informacji do trenowania, w celu ustalenia platformy (ang. framework) techniki, opartej na głębokim uczeniu stanowi empiryczne wyzwanie w obszarze medycyny. Zwiększa to ryzyko niewłaściwego wykorzystania danych osobowych. Bezpośrednim skutkiem tego problemu był gwałtowny rozwój badań nad metodami tworzenia syntetycznych danych pacjentów. Badacze podjęli próbę wygenerowania syntetycznych odczytów diagramów EKG lub PCG. Stąd, w celu zrównoważenia zbioru danych, w pierwszej kolejności utworzono dane EKG w bazie danych arytmii MIT-BIH przy użyciu struktur sieci generatywnych LSGAN i Cycle GAN. Następnie, wykorzystując strukturę sieci VGGNet, przeprowadzono badania, mające na celu klasyfikację arytmii na potrzeby syntetyzowanych sygnałów EKG. Dla wygenerowanych sygnałów, przypominających sygnał oryginalny uzyskano dobre rezultaty. Należy podkreślić, że uzyskana dokładność wynosiła 91,20%, powtarzalność 89,52% i wynik F1 – odpowiednio 90,35%.*

Słowa kluczowe: arytmia, osłuchiwanie, elektrokardiogram, fonokardiogram, sieci generatywne

Introduction

Cardiovascular diseases are a group of serious, life-threatening conditions that affect millions of people every year [23]. Cardiovascular disease is typically diagnosed in the clinic using information gathered from a variety of different detection methods. The disorders in heart beat or any abnormality can be termed as arrhythmias, which can range in severity from a minor inconvenience or pain to a life-threatening emergency. When the normal flow of the heart's electrical impulses is interrupted, an arrhythmia develops. It's possible that the heartbeat is too slow, too rapid, or otherwise irregular. Heart illness can be effectively predicted and diagnosed with the help of PCG and ECG signals. Artificial auscultation is a diagnostic tool for the heart that is efficient in terms of both time and money, but it requires clinicians to have extensive training in the field.

The electrocardiogram is a valuable tool for detecting arrhythmias or abnormalities as it enables the measurement of the bio-electrical activities of the heart whereas the PCG records the heart sounds. The bio-electrical activities of the heart can be recorded by means of an electrocardiogram (ECG), a noninvasive procedure. Electrodes attached to the patient's skin allowed an ECG machine to record the rhythmic contractions and relaxations of the heart. The T wave, the P wave, and the QRS complex are all components of a normal ECG signal. For example, Atrial fibrillation has been linked to abnormalities in the ECG, where there are no P-waves and an irregular ventricular rhythm [1]. Routinely, cardiologists undertake ECG screening for patients, which takes substantial human effort and costly medical procedures to detect heart problems and provide appropriate therapy for those concerns.

Regarding the PCG signals, medical professionals utilize a phonocardiograph to track the auditory manifestations and vibrations generated by the heart, encompassing those arising from the closure of the aortic, pulmonary, and atrioventricular valves. The information is depicted in figure 2. After the atrioventricular and semilunar valves close, the heart makes two distinct sounds, S1 and S2, during systole and diastole, respectively [3].

Due to the comparatively complicated waveform of the Phonocardiogram (PCG) signal in comparison to the Electrocardiogram (ECG), as well as the inherent challenges associated with the collecting approach that often result in significant disturbances, the widespread adoption and application of PCG signal analysis have been limited [16]. The reliability of visually assessing an electrocardiogram (ECG) for detecting heart abnormalities may be limited. The integration of deep network topologies into the automatic processing of ECG and PCG data, as well as other domains within the medical and healthcare industries, has become increasingly prevalent with the rise in popularity of deep learning approaches.

In recent years, the emergence of artificial intelligence as a potential option has been facilitated by technological advancements, particularly in the field of medical applications [7]. While deep learning algorithms are becoming increasingly used in e-health applications, large datasets are still required for the discovery of critical determinants that improve prediction or diagnostic accuracy.

The success of a deep learning model is highly dependent on the availability of both many training samples and a high-quality labelled dataset. Inadequate data or an imbalanced dataset might lead to a non-convergent training phase and biased classification results when training a deep learning model.

A big and well-balanced dataset is required for the Deep Learning model in order to avoid these problems. Despite having access to a big and evenly distributed dataset for training, would be ideal, this is not always possible because to factors such as the rarity of aberrant cardiac events and the scarcity of available cardiologists who can reliably identify (annotate) the waveforms. Because only highly trained doctors are capable of accurately or exactly annotating ECG recordings, the number of recordings that have been annotated is also restricted.

Therefore, there is a need for Deep learning algorithms to strive for the replication of novel artificial or synthetic data by initially comprehending the pattern derived from appropriate training data [8].

The Generative Adversarial Network (GAN) is a widely utilized form of data augmentation model that is employed for the creation of time series data as well as visualizations [10]. Various fields such as health care [22], stock market predictions [25], image segmentation [2], text classification [5] etc., have incorporated the utilisation of GANs.

In this research, we investigate how GANs can be used to generate authentic ECG and PCG signals from scratch. In Section II, we present an overview of relevant published materials. GAN, Least Squares GAN, and Cycle GAN architectures, as well as classification with VGGNet, are discussed in Section III; evaluation criteria are outlined in Section IV; and the simulated results and datasets are discussed in Section V.

1. Literature review

The conventional methods for the classification of electrocardiogram (ECG) and phonocardiogram (PCG) signals rely mostly on the use of traditional supervised machine learning approaches. Using a linear Support Vector Machine (SVM) [5], PCG signals were successfully classified. Numerous studies have used the SVM technique to classify electrocardiogram (ECG) readings [12].

These methods have not been able to scale adequately to incorporate ECGs from a large variety of individuals, even if we disregard the time and work necessary. Each patient's ECG signal will have its own unique dynamics and morphology. After much success in machine learning, researchers have started to take notice of the convolutional neural network's (CNN) potential for ECG and PCG categorization [15]. By coordinating the acquisition of ECG and phonocardiogram signals, Jan Nedoma et al. were able to make direct comparisons between the two methods of monitoring heart rate [19]. ECG signals in the time-frequency domain can be decomposed using wavelet techniques. Popular methods for feature extraction include principal component analysis and the hidden Markov model to name a few [14].

There are a number of deep learning methods that show promise for cardiovascular diseases classification and detection, but they need a lot of training instances to be truly effective. Due to the rarity of potentially lethal arrhythmias, there is a lack of data for training which is required for deep algorithms. There is a scarcity of deep PCG and ECG signals for rare diseases that are acceptable for clinical application because of the extensive heartbeat categorization challenge. To address this problem, we train deep learning models with simulated PCG and ECG signals depicting a variety of arrhythmias.

As a result, the need for medical image augmentation evolved. Basic data augmentation techniques encompass flipping, snipping, and introducing noise. In addition to these techniques, some other basic data augmentation techniques such as spatial inversion [11], time-spatial inversion [6], baseline wandering [20] are also applied to ECG signals. However, when it comes to the management of complex data such as medical imaging, these basic techniques are inadequate. The variational autoencoder (VAE) [20] is a deep model that has received less attention compared to other more widely adopted techniques. Nevertheless, a prominent concern revolves around the recurrent occurrence of hazy and indistinct output images.

2. Generative Adversarial Networks

In 2014, a group of researchers or academicians led by Ian Goodfellow [8] proposed the concept of GANs. Generative models, which include GANs, are a broad category. Therefore, GANs are based on the concept of zero-sum games, in which each player intentionally strives to maximise his or her own advantage at the expense of everyone else. The results of the GAN are a joint effort of a Generator and a Discriminator neural network. The Generator's goal is to perfect its ability to mislead the Discriminator with artificial distributions, whereas the Discriminator's goal is to perfect its capacity to recognise and reject fake distributions generated by the Generator. The figure 1 gives the architecture block diagram of GAN.

The GAN's effectiveness as a deep generative model can be attributed to its two main components, the generator and the discriminator. The generator (G) takes in a latent vector (z) with a Gaussian distribution and produces fake images or data. A discriminator trained on both forms of data x will produce an answer as to whether the created data is fake or real in its output (D). The goal function of a GAN is written out in the form of a min-max optimisation, as shown in equation 1.

In this study, we developed an LSGAN and Cycle GAN capable of producing realistic ECG and PCG signals, as well as a discriminator that can distinguish between the real and fake. Next these signals are given to a VGG-Net classifier to classify arrhythmias.

$$min_G max_D V(D,G) = \\ = E_Z \log\left(1 - D\left(G(z)\right)\right) + E_x \log\left(D(x)\right) \quad (1)$$

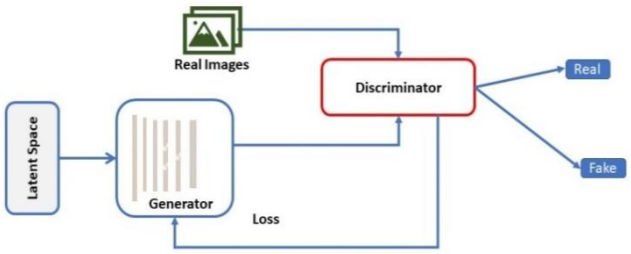


Fig. 1. Architecture of GAN

2.1. Cycle GAN

Cycle GANs can autonomously learn to translate between two visual inputs [27] when given just those two sets. Together, the two GANs in a Cycle GAN undergo training at the same time. The objective here is to preserve the constancy of the cycle at all times. In addition to the pair loss terms inherent in a GAN, a cycle loss term is included. It is necessary to optimise both GAN pairings in addition to the cycle loss term. The concept of "cycle loss" is visualised here. Take the task of teaching the Cycle GAN to convert summer (X) to winter (Y) landscapes as an example. To convert data from domain X to domain Y, we first train the initial generator, designated by F, to produce a winter image from a summer input image. DY is used to tell Y apart from the real thing. The second GAN pair does the opposite operation, converting X to Y and differentiating Y from X. While learning, the second GAN could invert Y into X. Therefore, while switching from summer to winter and back again, the images must remain visually consistent. This is shown in the figure 3 [27].

2.2. Least square GANs

GANs where the least-squares loss is utilised as a discriminator are called LSGANs [18]. Lowering the LSGAN's goal function has the same effect as minimising the Pearson 2 divergence. There are two benefits of LSGANs over classic GANs. First, unlike regular GANs, LSGANs can generate images of higher quality. Second, when it comes to teaching and learning, LSGANs are more reliable.

2.3. Arrhythmia classification

For the arrythmia classification, VGGNet architecture is used in this research. The VGGNet stands for Visual Geometry Group Network [21]. It was introduced in 2014, which was two years after AlexNet. The primary objective behind the development of this model was to investigate the influence that depth has on the level of precision achieved by picture classification training models. The model's speed and accuracy both saw substantial boosts after VGG was introduced. As the number of layers with smaller kernels was utilised, non-linearity improved, which is a desirable property in deep learning.

The VGG-Net is divided into VGG-16 and VGG-19 architectures. The VGG-16 network is a deep convolutional model that was trained using data from ImageNet. The major data set classification in this paper was accomplished using the VGG-16 model. Pooling layers, Convolution layers, and fully linked layers make up the network model. Figure4 depicts the network diagram of VGG-16's structural design. The VGG-16 architecture is shown in the figure 4.

The "16" in VGG stands for the number of layers in the deep neural network used by the VGGNet architecture. The VGG-16 architecture consists of 13 convolutional layers and three fully connected layers. The convolutional layer generates a feature map by applying a kernel matrix to the input matrix. By spreading the Kernel matrix over the input matrix, we are able to carry out a mathematical operation known as convolution. The feature map is the accumulated output of element-by-element matrix multiplication at each node. The output feature map of the convolutional layer remembers the exact position of the features in the input, which can be problematic. This means that the feature map will vary drastically if the input image is trimmed, or altered in any other way. As a solution, we employ down sampling in our convolutional layers. The ability to down sample can be achieved by introducing a pooling layer after a non-linearity layer. When the input is translated slightly, the representation becomes roughly invariant thanks to pooling.

In a convolutional neural network, the final Fully Connected Layer receives the output of the final Pooling Layer. There may be one, two, or all of these layers depending on the situation. To be fully connected, all of the first-layer nodes must have links to all of the second-layer nodes.

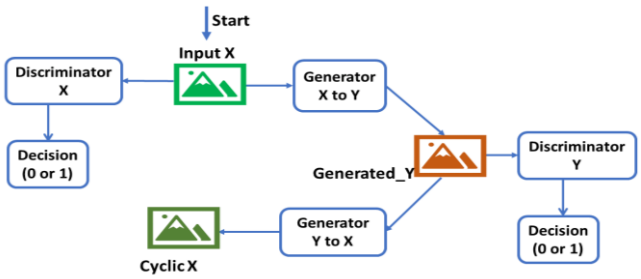


Fig. 2. Architecture of Cycle GAN

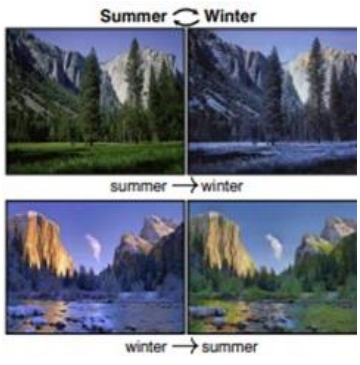


Fig. 3. Summer to winter [27]

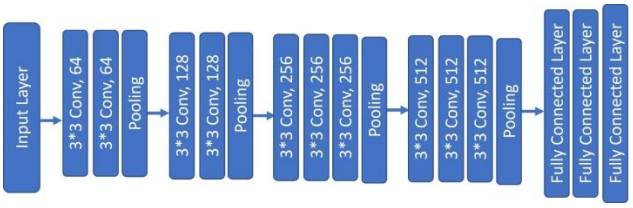


Fig. 4. Network Architecture of VGG-16

3. Evaluation metrics

Peak signal to noise ratio

The peak signal-to-noise ratio (PSNR) is the maximum pixel intensity divided by the distortion power, and it is calculated using the mean square error. Like MSE, the PSNR metric requires little effort to compute but may not correlate well with subjective evaluations of quality.

Structural similarity index

One metric for evaluating the quality of digital still photographs is the SSIM index. To calculate the SSIM of two images, x and y, we use the following formula. The Structural Similarity Index Approach is an interpretation-based framework. This approach sees image degradation as a shift in how we interpret the underlying structural details of an image. It works in tandem with other perceptual facts that are as crucial, such as the masking of brightness and contrast.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(c_1^2 + \mu_x^2 + \mu_y^2)(c_2^2 + \sigma_x^2 + \sigma_y^2)}$$

σ_y^2 – variance of y,
σ_x^2 – variance of x,
σ_{xy} – covariance of y and x,
μ_x – mean of x,
μ_y – mean of y.

Cross correlation coefficient

Cross-correlation assessment is a method of deducing properties of a signal from its correlation with another signal.

Accuracy

The degree of accuracy can be thought of as a representation of the probability that the values that were provided are accurate. A key measurement to take into account is the proportion of instances that the classifier correctly labelled.

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

TP indicates True Positive, FP indicates False Positive, FN indicates False Negative, and TN indicates True Negative.

Precision

The proportion of actual positive results relative to the total number of expected results is a measure of precision.

$$Precesion = \frac{TP}{TP + FP}$$

Recall

It evaluates a model based on how effectively it can provide accurate predictions.

$$Recall = \frac{TP}{TP + FN}$$

F1 score

It's a function that takes into account both accuracy and recall. It is the harmonic mean of recall and precision.

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

4. Results and discussion

4.1. Datasets

To evaluate the method's performance, we used two databases, PTB and MIT-BIH [9]. The PTB diagnostic database contains 549 records, representing 290 individuals. Each record contains data from fifteen consecutive measurements of the same signal. One thousand samples per second of digital data are taken from each signal. There is a wide variety of data available, including that pertaining to heart failure, myocardial hypertrophy, intra-cardiac, healthy persons, and more. There are 25 recordings with junctional, supra-ventricular, and heart block arrhythmias, and 23 healthy recordings with sequential numbers between 100 and 124 in the MIT-BIH database.

The PASCAL Heart Sound Challenge [4] dataset, the Heart Sound and Murmur Library [13], and the PhysioNet CinC Challenge dataset [17] are only a few examples of openly accessible datasets for PCG recordings.

Dataset 'A' and dataset 'B' are two subsets of the Pascal HSC database. Dataset 'A' was collected with the help of the iStethoscope Pro iPhone software, whereas dataset 'B' was compiled in a hospital setting with the help of the digital stethoscope. There is a total of 176 and 656, respectively, of auscultations in the two data sets. Both sets of data include recordings of regular heartbeats, as well as those with murmurs and additional systoles. The PhysioNet CinC database version 2016 was also created for a competition. The dataset includes heart sounds from clinical as well as non-clinical environments. The 'normal,' 'uncertain,' and 'abnormal' sound categories are all part of the test.

There is also the Heart sound and Murmur Library from the Michigan University Health Systems, which is a publicly available dataset containing many types of heart murmurs and sound, such as normal heart sounds, S1, single S2, split S2 transient, and more.

4.2. Simulation results

The real and artificial ECG and PCG signals generated by LS GAN are displayed in figure 5 and figure 6, respectively whereas figure 7 shows real and artificial ECG and PCG signals generated using Cycle GAN. The similarity results between real and artificial (or Generated) signals are shown in the table 1 whereas table 2 gives the arrhythmia classification results.

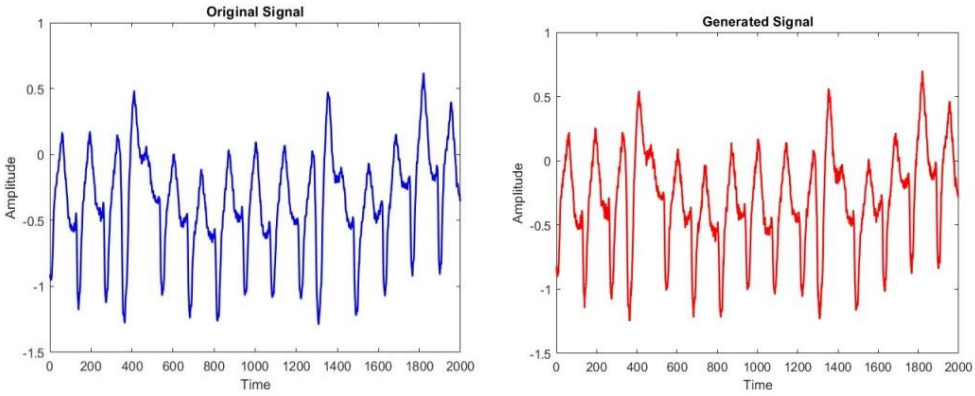


Fig. 5. Real and fake (generated) ECG Signals using LSGAN

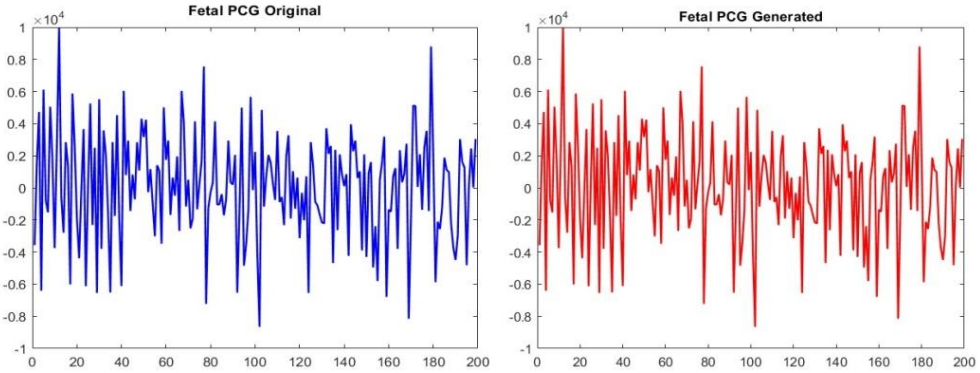


Fig. 6. Real and Fake (Generated) PCG Signals using LS GAN

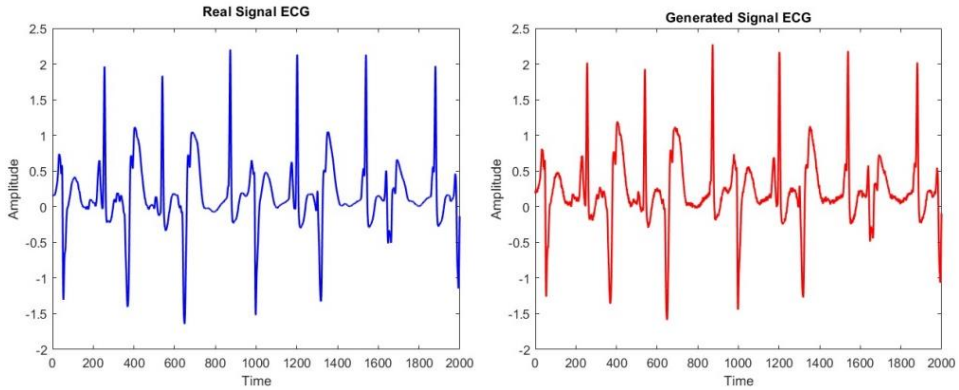


Fig. 7. Real and Fake (Generated) ECG Signals using Cycle-GAN

Table 1. Similarity results between synthesised and Real ECG and PCG signals

Method	Signal	MSE	SSIM
LSGAN	ECG	0.0702	0.9705
LSGAN	PCG	0.0728	0.9702
CycleGAN	ECG	0.0651	0.9823
CycleGAN	PCG	0.0699	0.9805

Table 2. Arrhythmia classification results

Method	Signal	Precision	Recall	F1 Score
LSGAN	ECG	0.9002	0.8569	0.8780
LSGAN	PCG	0.8959	0.8412	0.8676
CycleGAN	ECG	0.9120	0.8952	0.9035
CycleGAN	PCG	0.9054	0.8897	0.8974

5. Conclusion

There are a number of constraints that could make it difficult to get extensive patient information. The synthesis of realistic data has emerged as an attractive new topic of research in healthcare, particularly medicine. This is mostly due to the fact that it enables supervised machine learning classifiers to be better trained on datasets. In this research, ECG and PCG data were created using Least Squares GAN and Cycle GAN. Furthur, using VGGNet, studies were conducted to classify arrhythmias for the synthesized ECG signals. The synthesized signals performed well and resembled the original signal and the obtained precision of 96.99%, recall of 97.81% and an F1 score of 97.22%.

References

[1] Ahmed N., Zhu Y.: Early Detection of Atrial Fibrillation Based on ECG Signals. Bioengineering 7(1), 2020, 16 [http://doi.org/10.3390/bioengineering7010016].
[2] Akkaradamrongrat S. et al.: Text generation for imbalanced text classification. 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2019, 181–186.
[3] Aziz S. et al.: Phonocardiogram Signal Processing for Automatic Diagnosis of Congenital Heart Disorders through Fusion of Temporal and Cepstral Features. Sensors 20(13), 2020, 3790 [http://doi.org/10.3390/s20133790].
[4] Bentley P. et al.: Classifying Heart Sounds Challenge. 2011 [http://www.peterjbentley.com/heartchallenge/index.html]
[5] Bouril D. et al.: Automated classification of normal and abnormal heart sounds using support vector machines. Computing in Cardiology Conference – CinC, Vancouver 2016, 549–552.
[6] Cayce G. I. et al.: Improved Neural Network Arrhythmia Classification Through Integrated Data Augmentation. IEEE MetroCon, Hurst 2022, 1–3.
[7] England J. R., Cheng P. M.: Artificial intelligence for medical image analysis: a guide for authors and reviewers. American journal of roentgenology 212(3), 2019, 513–519.
[8] Garcea F. et al.: Data augmentation for medical imaging: A systematic literature review. Computers in Biology and Medicine 152, 2023, 106391 [http://doi.org/10.1016/j.compbiomed.2022.106391].
[9] Goldberger A. L. et al.: PhysioBank, PhysioToolkit and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101, 2000, 215–220.
[10] Goodfellow I. et al.: Generative adversarial networks. Communications of the ACM, 63(11), 2020, 139–144 [http://doi.org/10.1145/3422622].
[11] Guo G. et al.: Multimodal Emotion Recognition Using CNN-SVM with Data Augmentation. IEEE International Conference on Bioinformatics and Biomedicine, Las Vegas 2022, 3008–3014.
[12] Houssein E. H.: ECG signals classification: a review. International Journal of Intelligent Engineering Informatics 5(4), 2017, 376–396.
[13] Judge R., Mangrulkar R.: Heart Sound and Murmur Library. [http://open.umich.edu/education/med/resources/heart-sound-murmur-library/2015].
[14] Khalifa Y et al.: A review of Hidden Markov models and Recurrent Neural Networks for event detection and localization in biomedical signals. Information Fusion 69, 2021, 52–72.
[15] Li H. et al.: Dual-input neural network integrating feature extraction and deep learning for coronary artery disease detection using electrocardiogram and phonocardiogram. IEEE Access 7, 2019, 146457–146469.
[16] Li J., Ke L., Du Q., Ding X., Chen X.: Research on the Classification of ECG and PCG Signals Based on BiLSTM-GoogLeNet-DS. Applied Sciences 12(22), 2022, 11762.
[17] Liu C. et al.: An open access database for the evaluation of heart sound algorithms. Physiological Measurement 37(12), 2016, 2181.
[18] Mao X. et al.: Least Squares Generative Adversarial Networks. arXiv, 2017 [http://arxiv.org/abs/1611.04076].
[19] Nedoma J. et al.: Comparison of BCG, PCG and ECG signals in application of heart rate monitoring of the human body. 40th International Conference on Telecommunications and Signal Processing – TSP, 2017, 420–424.
[20] Rahman, M. M. et al.: A Systematic Survey of Data Augmentation of ECG Signals for AI Applications. Sensors, 23(11), 2023, 5237 [http://doi.org/10.3390/s23115237].
[21] Simonyan K., Zisserman A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv, 2015 [http://arxiv.org/abs/1409.1556].
[22] Skandarani Y. et al.: GANs for medical image synthesis: An empirical study. Journal of Imaging 9(3), 2023 [http://doi.org/10.3390/jimaging9030069].
[23] Sreeniwas Kumar A., Nakul S.: Cardiovascular Disease in India: A 360 Degree Overview. Medical Journal Armed Forces India 76(1), 2020, 1–3 [http://doi.org/10.1016/j.mjafi.2019.12.005].
[24] Wang T. C. et al.: High-resolution image synthesis and semantic manipulation with conditional gans. IEEE Conference on computer vision and pattern recognition. Salt Lake City, 2018, 8798–8807.
[25] Wu J. L. et al.: A prediction model of stock market trading actions using generative adversarial network and piecewise linear representation approaches. Soft Comput 27, 2023, 8209–8222 [http://doi.org/10.1007/s00500-022-07716-2].
[26] Xiong P. et al.: Deep Learning for Detecting and Locating Myocardial Infarction by Electrocardiogram: A Literature Review. Frontiers in Cardiovascular Medicine 9, 2022 [http://doi.org/10.3389/fcvm.2022.860032].
[27] Zhu J. Y. et al.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. arXiv [http://arxiv.org/abs/1703.10593].

M.Sc. Swarajya Madhuri Rayavarapu
e-mail: madhurirayavarapu.rs@andhrauniversity.edu.in

Currently pursuing Ph.D. in Department of Electronics and Communication, Andhra University. She obtained her M.Tech Degree from CASEST, University of Hyderabad.
Her research interests include Deep Learning, Generative Adversarial Networks (Semi-supervised Machine Learning) in medical Image Processing, Applying deep learning techniques to 5G-Mobile Communication (Layer 2 of RAN).

http://orcid.org/0009-0007-7559-2142

M.Sc. Tammineni Shanmukha Prasanthi
e-mail: prashanthitammineni.rs@andhrauniversity.edu.in

Tammineni Shanmukhaprasanthi obtained M.Tech. degree in Electronics and Communication Engineering from Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh in 2021. She is currently pursuing Ph.D. degree with Andhra University Visakhapatnam, India. Her research interests include microstrip patch antenna design, VLSI circuit design, image inpainting in image processing.

http://orcid.org/0009-0000-5352-2265

Ph.D. Gottapu Santosh Kumar
e-mail: kumar.santou@gmail.com

He is assistant professor in the Department of Civil Engineering, Gayatri Vidya Parishad College of Engineering. He completed his Ph.D. degree from Andhra University College of Engineering, Visakhapatnam, India. His research interests include Artificial Intelligence, machine learning.

http://orcid.org/0000-0002-1452-9752

Prof. Gottapu Sasibhushana Rao
e-mail: sasigps@gmail.com

He is Senior Professor in the Department of Electronics & Communication Engineering, Andhra University College of Engineering, Visakhapatnam, India. He is a senior member of IEEE, fellow of IETE, member of IEEE communication Society, Indian Geophysical Union and International Global Navigation Satellite System, Australia. Prof. Rao was also the Indian member in the International Civil Aviation organization, Canada working group for developing SARPS.

http://orcid.org/0000-0001-6346-8274

Ph.D. Gottapu Prashanti
e-mail: prashanti.gottapu@gmail.com

She is Assistant professor in the Department of Pharmaceutics Technology, Avanthi Cavanthi Institute of Pharmaceutical Sciences. She completed her Ph.D. degree from Andhra University, Visakhapatnam, India. Her research interests include Artificial Intelligence, machine learning.

http://orcid.org/0009-0002-7231-0377

http://doi.org/10.35784/iapgos.5332 *received: 20.09.2023 | revised: 22.10.2023 | accepted: 25.11.2023 | available online: 20.12.2023*

SMART OPTIMIZER SELECTION TECHNIQUE: A COMPARATIVE STUDY OF MODIFIED DENSNET201 WITH OTHER DEEP LEARNING MODELS

Kamaran H. Manguri[1,2], Aree A. Mohammed[3]
[1]Erbil Polytechnic University, Erbil Technical Engineering College, Department of Technical Information System Engineering, Erbil, Iraq, [2]University of Raparin, Department of Software and Informatics Engineering, Erbil, Iraq, [3]University of Sulaimani, College of Science, Computer Science Department, Sulaymaniyah, Iraq

Abstract. *The rapid growth and development of AI-based applications introduce a wide range of deep and transfer learning model architectures. Selecting an optimal optimizer is still challenging to improve any classification type's performance efficiency and accuracy. This paper proposes an intelligent optimizer selection technique using a new search algorithm to overcome this difficulty. A dataset used in this work was collected and customized for controlling and monitoring roads, especially when emergency vehicles are approaching. In this regard, several deep and transfer learning models have been compared for accurate detection and classification. Furthermore, DenseNet201 layers are frizzed to choose the perfect optimizer. The main goal is to improve the performance accuracy of emergency car classification by performing the test of various optimization methods, including (Adam, Adamax, Nadam, and RMSprob). The evaluation metrics utilized for the model's comparison with other deep learning techniques are based on classification accuracy, precision, recall, and F1-Score. Test results show that the proposed selection-based optimizer increased classification accuracy and reached 98.84%.*

Keywords: deep learning, optimization technique, transfer learning, customized dataset, modified DenseNet201

INTELIGENTNA TECHNIKA WYBORU OPTYMALIZATORA: BADANIE PORÓWNAWCZE ZMODYFIKOWANEGO MODELU DENSENET201 Z INNYMI MODELAMI GŁĘBOKIEGO UCZENIA

Streszczenie. *Szybki wzrost i rozwój aplikacji opartych na sztucznej inteligencji wprowadzają szeroki zakres architektur modeli głębokiego uczenia i uczenia transferowego. Wybór optymalnego optymalizatora wciąż stanowi wyzwanie w celu poprawy wydajności i dokładności każdego rodzaju klasyfikacji. W niniejszej pracy proponowana jest inteligentna technika wyboru optymalizatora, wykorzystująca nowy algorytm wyszukiwania, aby pokonać to wyzwanie. Zbiór danych użyty w tej pracy został zebrany i dostosowany do celów kontroli i monitorowania dróg, zwłaszcza w sytuacjach, gdy zbliżają się pojazdy ratunkowe. W tym kontekście porównano kilka modeli głębokiego uczenia i uczenia transferowego w celu dokładnej detekcji i klasyfikacji. Ponadto, warstwy DenseNet201 zostały zamrożone, aby wybrać optymalizatora idealnego. Głównym celem jest poprawa dokładności klasyfikacji samochodów ratunkowych poprzez przeprowadzenie testów różnych metod optymalizacji, w tym (Adam, Adamax, Nadam i RMSprob). Metryki oceny wykorzystane do porównania modelu z innymi technikami głębokiego uczenia opierają się na dokładności klasyfikacji, precyzji, czułości i miarze F1. Wyniki testów pokazują, że zaproponowany optymalizator oparty na wyborze zwiększył dokładność klasyfikacji i osiągnął wynik na poziomie 98,84%.*

Słowa kluczowe: głębokie uczenie, technika optymalizacji, uczenie transferowe, dostosowany zbiór danych, zmodyfikowany DenseNet201

Introduction

Due to the rapid population growth in the world, the number of cars and usage of vehicles have exponentially risen [4], which has led to increasingly congested roads, heightened air pollution, and a rise in accidents [10]. In such a scenario, effective traffic monitoring of the streets is required, a severe challenge in many cities worldwide. Also, urban traffic congestion can potentially result in various impacts on the environment, public health, and the economic situation. Consequently, an accurate road transportation system is required based on Traffic Signal Timing (TST), considered an optimized and fastest technique to reduce the congestion at the road's intersections and improve the urban traffic flow networks [19]. The main challenge in intelligent transportation focuses on controlling traffic signals, impacting the provision of transportation services within urban transportation systems [18]. Besides, intelligent transportation systems can be utilized to overcome the related issues to traffic signals. In particular, the problems that constantly affect our society, namely (human safety, waiting time minimization, effective cost reduction, and decrease of carbon dioxide emission (CO2) to the atmosphere.

Moreover, launching signal timing control ensures individuals' safe and efficient movement through an intersection. Achieving this goal necessitates a well-structured strategy for determining the right-of-way allocation. To make the optimization task attainable, it is necessary to introduce specific assumptions. However, a notable challenge arises from frequent discrepancies and occasionally substantial divergence between these assumptions and real-world conditions. In the meantime, numerous factors influence drivers within real-world traffic scenarios, including interactions influenced by driver preferences with vulnerable road users (such as pedestrians, cyclists, etc.), weather conditions, and road quality [7]. Emergency vehicles play a critical role in various life-threatening situations. Over 20% of patients in ambulances are at risk due to traffic congestion. However, the mortality rate tends to increase in severely ill patients. Countries with high population densities often face significant traffic congestion during peak hours. Emergency vehicles like police, ambulances, and firefighters frequently become trapped in traffic, leading to life-threatening circumstances. Therefore, prioritizing these emergency vehicles and ensuring their seamless flow becomes essential. This goal can be achieved by suggesting an automated traffic system to identify these vehicles even when roads are congested [21].

Additionally, accurate vehicle detection systems are essential for distinguishing between emergency and regular vehicles. Deep learning and transfer learning approaches employing Convolutional Neural Networks (CNNs) have been harnessed in computer vision, producing results that match human capabilities and even surpass those of human experts. The Deep convolutional neural network is one of the best techniques for detecting and classifying objects in image and video [17]. The general concept uses knowledge learned from the pre-trained neural network model applied to different but related problems. Various pre-trained models including AlexNet, Visual Geometry Group (VGG)-19, ResNet-101, inception-v3, and DenseNet201 [14] are applied to the customized dataset of non-emergency and emergency cars.

Most recent advances and developments for deep learning applications in many fields have been presented, such as computer vision, medical image, natural language processing, speech processing, and traffic congestion [5]. A modified version of MobileNet based on the increase in frame rate is proposed by Ahmed et al. to achieve higher F-values. It could be applied to diverse applications supporting real-time traffic data analysis [2]. Another deep learning model based on ResNet-50 is presented for vehicle localization and classification using real data from traffic surveillance cameras [12]. Automatic damage vehicle classification has become an interesting area for researchers. An image-based inspection model using an adapted version

of pre-trained convolutional neural networks namely VGG-19 and VGG-16 is proposed to reduce the damage to car's cost [6].

In [1] a new deep-learning model for Bangladesh vehicle classification is proposed using a Multi-class Vehicle Image Network (MVINet) based on DenseNet201 with four additional layers. The proposed approach utilized CNNs to extract features from the image vehicles dataset.

This paper uses different optimizers to improve the efficiency and accuracy of the proposed deep and transfer models. The optimization process undertakes multiple paths until reaching the convergence. Various optimization mechanisms have arisen to solve the problems during the learning process. The following optimizers are taken into consideration including the Stochastic Gradient Descent (SGD), Stochastic Optimization Descent with Momentum, Rung Kutta, Adaptive Learning Rate, Root Mean Square Propagation, Adaptive Moment Estimation, Deep Ensembles, Feedback Alignment, Direct Feedback Alignment, Adfactor, AMSGrad, and Gravity. Experimental results illustrated that best optimizers like SGD and Adam have optimal training, testing, and validation accuracy [8]. The general block diagram of the proposed system is depicted in Fig. 1.

This study applied various deep learning techniques to classify vehicles while passing in a traffic intersection. The paper's organization is as follows: Section I provides a general introduction and its associated issues. Section II entails the background of traffic management. Section II presents relevant prior research and surveys previous studies. The methodology is explained in Section III. In Section IV, results are shown and discussed. The concluding remarks are provided in the final section.

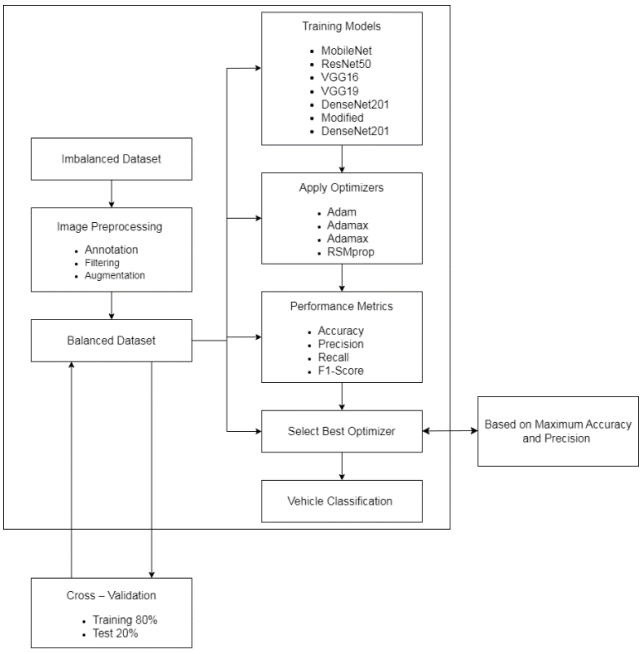


Fig. 1. General diagram of proposed technique

1. Motivation and contribution

In the Kurdistan Region of Iraq (KRG), considered a developing region in Iraq, the main challenges facing people are accidents at the intersection of traffic signaling and control systems. Many factors are involved in such an unreliable environment, such as (planning urban roads, lack of driver awareness, and lack of traffic signs). Furthermore, emergency vehicles require an efficient traffic system to reduce road congestion. To build this system, which is the main aim of this research, a customized dataset is created for ambulance, police, and firefighter cars.

Recently, researchers have mainly used deep learning techniques for emergency vehicle detection and classification based on CNN and RNN architectures. Nevertheless, few researchers use the DenseNet201 transfer model to address the abovementioned issues. The crucial contributions are summarized as follows:

- The proposed models use different image sizes (64*64, 128*128, and 224*224) as an input for the model's training. Various transformations also augment the images to make the emergency vehicles dataset balanced.
- Applying a novel search algorithm, which is based on performance metrics namely (accuracy and precision) to select the best optimizer that has been tested with various deep and transfer learning models such as (MobileNet, VGG-16, VGG-19, ResNet, DenseNet-201, etc.).
- Freezing DenseNet-201 layers to enhance performance accuracy of the vehicle's detection and classification.

2. Related works

Researchers have done much work to detect and classify vehicles in traffic intersections and use the acquired data to control traffic signal systems. Deep learning techniques are mainly used to improve those system accuracies for emergency vehicles regarding detection and classification, while few studies have utilized transfer learning-based models. They depend on their proposed techniques' image or video preprocessing and tuning hyper-parameters. Kamaran H. and Aree A. [16] extensively surveyed computer vision-based traffic control and monitoring. The aim was to comprehensively review recent research using computer vision and deep learning algorithms. They also compared the models based on the algorithms, dataset, and performance accuracy used, respectively. In addition, Tomar et al. [22] presented a state-of-the-art review of smart vehicles, including challenges and emerging trends. They proposed solutions for the transportation system, especially controlling and managing traffic at the road intersections. On the one hand, they included synchronizing signals across a specific route and improving the real-time traffic flow at junctions and intersections while cars and emergency vehicles pass. On the other hand, achieved results indicated a remarkable reduction in congestion by applying signal synchronization to the busy roads in the smart cities. In [15], Dallas et al. offered a comprehensive overview aimed at improving the traffic signal system's safety and accuracy through various performance evaluation metrics. Their survey also had a potential benefit to traffic agencies, researchers, and commercial offices. The simulated results performed by several methodologies showed that when sensor data and probe vehicles are used, the performance accuracy of the traffic signal system is better.

To incorporate the machine and deep learning models in traffic flow prediction, Noor et al. [20] introduced a critical review to present the techniques and gaps in the Intelligent Transportation System (ITS) performance evaluation. Based on their findings, the most frequent ML techniques for predicting traffic flow are CNN-based and Long-Short Memory (LSTM).

Joo has proposed an improved reinforcement learning model for traffic signaling management systems, H. et al. [11]. They tackled the traffic congestion problem using an effective Traffic Signal Control (TSC) algorithm, which adaptively increases the number of normal and emergency cars passing through an intersection and balances the signals between roads using Q-learning (QL). Based on their simulated results, the proposed model performs better than other research using (QL). Ke, X. et al. [13] designed and developed a model for detecting road congestion in ITS based on CNN and multidimensional visual features. Firstly, the model detects the density estimation of foreground objects utilizing a gray-level co-occurrence matrix; then, the Lucas–Kanade optical flow algorithm with a pyramid implementation is performed to measure the speed of objects in movement. Secondly, a Gaussian mixture method is used for the background estimation, and CNN accurately separates the final foreground from the foreground candidates. Finally, the attained results through objective and subjective metrics

are evaluated to exceed the state-of-the-art research work in road-traffic congestion detection methods. Currently, roads and highways in urban networks are supplied with many surveillance cameras that can be used for real-time vehicle detection and identification by efficiently estimating traffic flow. The paper presented a comparative study of object detectors, visually described features, and classification models to implement traffic estimations. Notably, three object detectors are used to categorize vehicles. Additionally, four traditional machine learning models are sequentially employed to extract visual features for classification. These traditional ML approaches are compared to the deep learning-based models. The research findings indicate that when techniques are accordingly implemented and tested, the deep learning method performs most accurately, especially for multi-class classification [9].

Ashir M. introduced a real-time detection method for various emergency vehicles based on a computer vision approach under massive traffic conditions. This allows priority control to emergency vehicles by the traffic controller agencies, which significantly saves people's lives and properties, prevents crimes, and forcefully minimizes the time needed by an emergency vehicle to arrive at its destination. A modified version of the YOLOv5 algorithm for object detection is proposed and applied to four types of emergency vehicles. YOLO is considered one of the most well-known object detection schemes due to its efficiency and accuracy. The connected layers of the YOLO algorithm are modified to have new learning weights while freezing convolutional layers. The test results of the proposed model have indicated promising results for emergency vehicle detection and classification [3].

The abovementioned survey states that conventional deep learning techniques cannot accurately obtain an optimal result. Hence, a transfer learning-based model, DenseNet-201, is developed to fill this gap with some improvements. It is dynamically adapted to select one of the best optimizers used in this research work.

3. Proposed methods

In this study, four optimizers are tested with different deep and modified transfer learning techniques using a customized dataset created by [17]. As shown in figure 1, the proposed methodology steps are as follows:

- Dataset collection: Emergency and non-emergency vehicles are gathered from a publicly available dataset and local traffic offices in KRG- Iraq. The vehicle class types are imbalanced.
- Image labeling: It is based on emergency (Ambulance, Police, and Firefighter) and non-emergency cars.
- Image quality enchantment: perform preprocessing algorithms, including image resizing, sharpening, smoothing, and contrast enhancement.
- Data augmentation: to balance the dataset, various image transformations are applied to overcome the overfitting issue.
- Dataset splitting: Datasets are portioning to cross-validation of 80% (training data) and 20% (test and validation).
- Train the proposed deep and transfer models with tuning hyper-parameters of each model.
- Use the proposed search mechanism to find out the best optimizer. The condition is based on returning the maximum value of the accuracy and precision metrics.
- Performance metrics' evaluation: precision, recall, f1-score, and average accuracy.
- Vehicles classification: The multi-classification method identifies emergency and non-emergency vehicles.

3.1. Customized dataset

Because no publicly available dataset contains emergency cars (police, ambulance, and firefighters) and non-emergency cars, a customized dataset has been produced. In this work, the used dataset was collected from the (Kaggle, Fatkun Batch, and Rania traffic directorate in Kurdistan reign–Iraq). In table 1, the dataset for both balanced and unbalanced datasets is displayed.

3.2. Modified DenseNet201 model

In this study, a modified version of "Densely Connected Convolutional Neural Network" (DenseNet201) is developed to improve and overcome the state-of-art, which is based on traditional CNN for feature extraction and classification.

Table 1. Balanced and Unbalanced Datasets

Vehicle types	Unbalanced data size	Balanced data size
Ambulance	322	1610
Firefighters	526	1682
Police car	700	1260
Non-emergency	1670	1670

Concerning neural networks, freezing layers can be considered a mechanism in which the control of learning weight parameters is updated. When the layers are frozen, the weights cannot be further modified during potential processing. This strategy is to reduce the computational complexity for training, whereas it does not have a significant effect on detection accuracy. Therefore, freezing the model's layers is a method to speed up the neural network training by freezing hidden layers in progressive ways. In this regard, many freezing layers are performed on DenseNet201 layers in the interest of its impact on performance accuracy. Figure 2 demonstrates the modified architecture of DenseNet201 by freezing some layers.

3.3. Proposed Search Algorithm

All models trained with the input size of (64*64, 128*128, 224*224) are simulated for different epoch numbers. For each model, the performance metrics are calculated based on accuracy, precision, recall, and f1-score, respectively. The conducted results are then sorted to find out the maximum values of accuracy and precision. Our search mechanism relies on the condition where the selection of the best optimizer is performed when both accuracy and precision measures have maximum values, as depicted in Fig. 3.

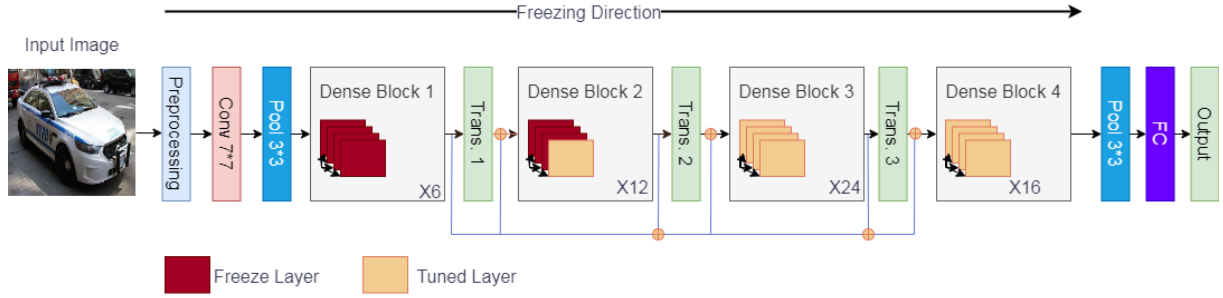


Fig. 2. Modified DenseNet201 architecture

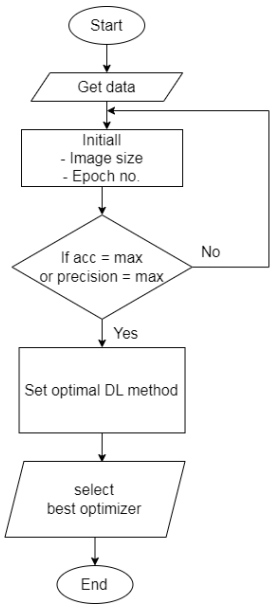


Fig. 3. Proposed Search Mechanism

3.4. Performance evaluation metrics

The proposed models are trained with the input mentioned above data for a range of iterations. Performance metrics such as accuracy, precision, recall, and f1-scores are used for estimating the model's performance accuracy. The following equations specify these metrics accordingly [23].

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

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

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

$$\text{F1} = \frac{2\ x\ Recall\ x\ Precision}{Recall+\ Precision} \quad (4)$$

Where true positive, true negative, false positive, and false negative are expressed by TP, TN, FP, and FN respectively.

4. Results and discussions

As explained in previous sections, the main goal of this work is to find an optimal optimizer that can effectively comply with proposed deep and transfer learning techniques. Accordingly, the pre-trained models are compiled using the customized dataset for different epoch numbers (i.e., iterations). Other hyper-parameters are fixed for all methods, such as (filter size=7*7, dropout_rate = 0.5, batch_size = 16, and learning-rate = 0.0001). The conducted tests determine the performance metrics (Average accuracy, precision, recall, and F1-Score) over 15-30 iterations. Fig. 4 and 5 show the accuracy and precision results for the different input image sizes when the best optimizer has been chosen. In this case, a modified DenseNet201 with 120 layers freezing is used.

Experimental results in Figures 4 and 5 reveal that the image size 224*224 gives better accuracy and precision values. Therefore, for further tests, we merely used this size with all techniques and optimizers in this research.

Table 2 and 3 present the accuracy and precision metrics results to select the best optimizer using our new search strategy.

The highlighted rows in the above tables indicate that the more accurate model is the DenseNet201 with 120 layers freezing and the best optimizer is RMSprop for both accuracy and precision tests when the input image size is 224*224 and the epoch No. is 30.

In addition, the confusion matrix of the proposed emergency vehicles classification transfer-based model is plotted, which is displayed in Fig. 6. The matrix diagonal represents the proposed model performance accuracy for different types of emergency cars.

The model's (DenseNet201 with 120 layers freezing) loss and accuracy for the training and test validation are depicted in Fig. 7 and Fig. 8 respectively.

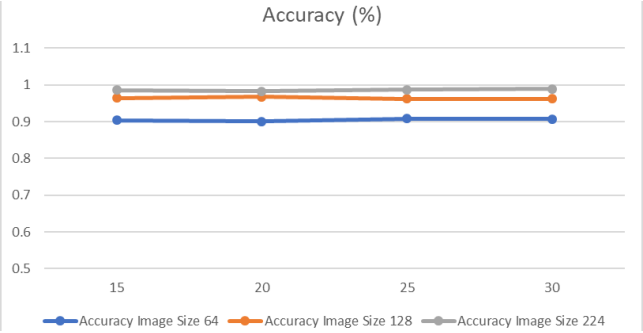


Fig. 4. Accuracy versus image size (DensNet201-120_freeze_layers)

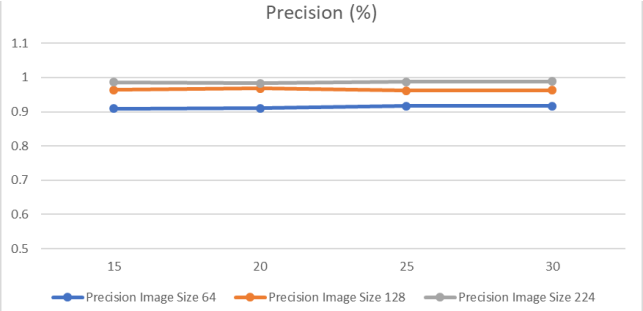


Fig. 5. Precision versus image size (DensNet201-120_freeze_layers)

Table 2. Accuracy tests

Image Size	Epoch No.	Models	Optimizers	Accuracy (%)
64	15	VGG19	Adam	90.37
	20	VGG19	Adam	90.04
	25	VGG16	RMSprop	90.78
	30	VGG19	Nadam	90.69
128	15	DenseNet201 Freeze 0	Nadam	96.37
	20	DenseNet201 Freeze 0	RMSprop	96.79
	25	DenseNet201 Freeze 0 - 30	Adam	96.21
	30	DenseNet201 Freeze 0 - 90	Adam	96.21
224	15	DenseNet201 Freeze 0 - 30	Adam	98.06
	20	DenseNet201 Freeze 0 - 60	RMSprop	98.27
	25	DenseNet201 Freeze 0 - 150	RMSprop	98.68
	30	DenseNet201 Freeze 0 - 120	RMSprop	98.84

Table 3. Precision tests

Image Size	Epoch No.	Models	Optimizers	Accuracy (%)
64	15	VGG16	Adam	90.96
	20	VGG19	Adam	91.01
	25	VGG16	RMSprop	91.63
	30	VGG19	Nadam	91.65
128	15	DenseNet201 Freeze 0	Nadam	96.37
	20	DenseNet201 Freeze 0	RMSprop	96.84
	25	DenseNet201 Freeze 0 - 30	Adam	96.21
	30	DenseNet201 Freeze 0 - 90	Adam	96.26
224	15	DenseNet201 Freeze 0 - 30	Adam	98.61
	20	DenseNet201 Freeze 0 - 60	RMSprop	98.27
	25	DenseNet201 Freeze 0 - 150	RMSprop	98.71
	30	DenseNet201 Freeze 0 - 120	RMSprop	98.85

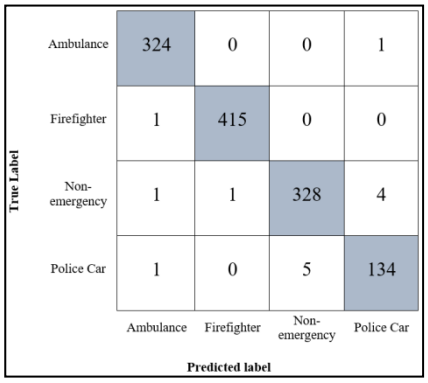


Fig. 6. Confusion matrix for the optimal model and optimizer

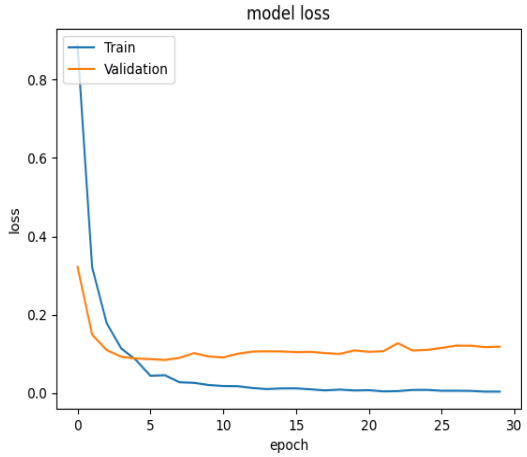


Fig. 7. Model's loss (DensNet201 – 120 layers freezing)

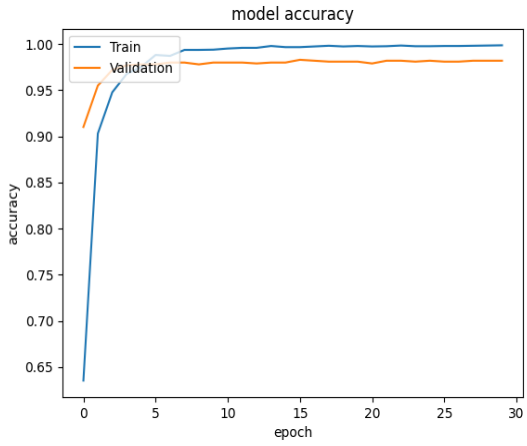


Fig. 8. Model's accuracy (DensNet201 – 120 layers freezing)

5. Conclusions

This paper uses a new search mechanism to develop an enhanced deep transfer learning model based on freezing layers. It is designed for emergency and non-emergency vehicle detection and classification. Various optimizers are used for training different models to find out the optimal accuracy of detection. In the initial phase, various image preprocessing and augmentation methods are employed on the input image dataset to achieve dataset balance. Afterward, the basic architecture of the DensNet201 model is modified by freezing the layers. Finally, the trained models are simulated with various optimizers to select the optimal one. The higher accuracy attained by the proposed method is 98.84% when the RMSProp is used. To reduce congestion and waiting time, the future work plan will test the proposed model in a real-time traffic signaling system when different emergency vehicles enter the lane intersections.

Acknowledgments

The authors wish to thank the University of Sulaimani, Erbil Polytechnic University, and the AI Research Lab at the University of Raparin for providing a scientific environment to complete this work.

Conflicts of interest

The author certifies that there is no actual or potential conflict of interest concerning this article.

Availability statement

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

[1] Ahmed T. et al.: A Deep Learning based Bangladeshi Vehicle Classification using Fine-Tuned Multi-class Vehicle Image Network (MVINet) Model. 2023 International Conference on Next-Generation Computing, IoT and Machine Learning – NCIM, 2023, 1–6.
[2] Ahmed U. et al.: Multi-aspect detection and classification with multi-feed dynamic frame skipping in vehicle of internet things. Wireless Netw, 2022, 1–12.
[3] Ashir S. M. et al.: A Transfer-Learning-Based Approach for Emergency Vehicle Detection. Eurasian Journal of Science and Engineering 8(1), 2022.
[4] Biswas D. et al.: An automatic car counting system using OverFeat framework. Sensors 17(7), 2017, 1535.
[5] Dong S. et al.: A survey on deep learning and its applications, Computer Science Review 40, 2021, 100379.
[6] Fouad M. M. et al.: Automated vehicle inspection model using a deep learning approach. J Ambient Intell Human Comput 14, 2023, 13971–13979.
[7] Ghazal B. et al.: Smart traffic light control system. Third international conference on electrical, electronics, computer engineering and their applications – EECEA, 2016, 140–145.
[8] Hassan E. et al.: The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study. Elmougy and Applications 82(11), 2023, 16591–16633.
[9] Impedovo D. et al.: Vehicular traffic congestion classification by visual features and deep learning approaches: a comparison. Sensors 19(23), 2019, 5213.
[10] Jain N. K. et al.: A review on traffic monitoring system techniques. SoCTA 2019, 569–577.
[11] Joo H. et al.: Traffic signal control for smart cities using reinforcement learning. Computer Communications 154, 2020, 324–330.
[12] Jung H. et al.: ResNet-based vehicle classification and localization in traffic surveillance systems. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, 61–67.
[13] Ke X. et al.: Multi-dimensional traffic congestion detection based on fusion of visual features and convolutional neural network. IEEE Transactions on Intelligent Transportation Systems 20(6), 2018, 2157–2170.
[14] Khan A. et al.: A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 53, 2020, 5455–5516.
[15] Leitner D. et al.: Recent advances in traffic signal performance evaluation. Journal of Traffic and Transportation Engineering 9(4), 2022, 507–531.
[16] Manguri K. H. K. et al.: A Review of Computer Vision–Based Traffic Controlling and Monitoring. UHD Journal of Science and Technology 7(2), 2023, 6–15.
[17] Manguri K. H. K., Mohammed A. A: Emergency vehicles classification for traffic signal system using optimized transfer DenseNet201 model. Indonesian Journal of Electrical Engineering and Computer Science 32(2), 2023, 1058–1068.
[18] Mohammad M. A. et al.: New Ontology structure for intelligent controlling of traffic signals. Procedia Computer Science 207, 2022, 1201–1211.
[19] Qadri S. S. S. M. et al.: State-of-art review of traffic signal control methods: challenges and opportunities. Eur. Transp. Res. Rev. 12(55), 2020, 1–23.
[20] Razali N. A. M. et al.: Gap, techniques and evaluation: traffic flow prediction using machine learning and deep learning. J Big Data 8(1), 2021, 1–25.
[21] Roy S., Rahman M. S.: Emergency vehicle detection on heavy traffic road from cctv footage using deep convolutional neural network. International Conference on Electrical, Computer and Communication Engineering – ECCE, 2019, 1–6.
[22] Tomar I. et al.: State-of-Art review of traffic light synchronization for intelligent vehicles: current status, challenges, and emerging trends. Electronics 11(3), 2022, 465.

M.Sc. Kamaran Manguri
e-mail: kamaran@uor.edu.krd

Lecture in the Department of Computer Science, College of Basic Education, University of Raparin. He had received Master Degree in Electronics and Computer Engineering, Hasan Kalyoncu University, Turkey, and now he is a Ph.D. student in Erbil Technical Engineering College, Erbil Polytechnic University Major field (Computer vision and Machine Learning).

http://orcid.org/0000-0001-8567-3367

Prof. Aree Ali Mohammed
e-mail: kamaran@uor.edu.krd

Born in Sulaimani city–Kurdistan Region Iraq. He obtained a B.Sc. degree at the University of Mousle (1995), an M.Sc. degree in France in Computer Science (2003), and a Ph.D. In multimedia systems at the Univesity of Sulaimani (2008). He directed Information Technology Directorate for four years (2010– 2014) and the head of the Computer Science Department / College of Science / University of Sulaimani for seven years. The main field of interest is multimedia system applications for processing, compression, and security. Many papers have been published in scientific journals throughout the world.

http://orcid.org/0000-0001-9710-4559

http://doi.org/10.35784/iapgos.5392 *received: 05.08.2023 | revised: 14.10.2023 | accepted: 25.11.2023 | available online: 20.12.2023*

IMPROVEMENT OF THE ALGORITHM FOR SETTING THE CHARACTERISTICS OF INTERPOLATION MONOTONE CURVE

Yuliia Kholodniak[1], Yevhen Havrylenko[2], Serhii Halko[3], Volodymyr Hnatushenko[4], Olena Suprun[5], Tatiana Volina[6], Oleksandr Miroshnyk[7], Taras Shchur[8]

[1]Dmytro Motornyi Tavria State Agrotechnological University, Department of Computer Sciences, Zaporizhzhia, Ukraine, [2]Dmytro Motornyi Tavria State Agrotechnological University, Department of Engineering Mechanics and Computer Design, Zaporizhzhia, Ukraine, [3]Dmytro Motornyi Tavria State Agrotechnological University, Department of Electrical Engineering and Electromechanics named after Prof. V.V. Ovharov, Zaporizhzhia, Ukraine, [4]Dnipro University of Technology, Department of Information Technologies and Computer Engineering, Dnipro, Ukraine, [5]Dmytro Motornyi Tavria State Agrotechnological University, Department of Foreign Languages, Zaporizhzhia, Ukraine, [6]National University of Life and Environmental Sciences of Ukraine, Department of Descriptive Geometry, Computer Graphics and Design, Kiev, Ukraine, [7]State Biotechnological University, Department of Electricity Supply and Energy Management, Kharkiv, Ukraine, [8]Cyclone Manufacturing Inc, Mississauga, Ontario, Canada

Abstract. *Interpolation of a point series is a necessary step in solving such problems as building graphs de-scribing phenomena or processes, as well as modelling based on a set of reference points of the line frames defining the surface. To obtain an adequate model, the following conditions are imposed upon the interpolating curve: a minimum number of singular points (kinking points, inflection points or points of extreme curvature) and a regular curvature change along the curve. The aim of the work is to develop the algorithm for assigning characteristics (position of normals and curvature value) to the interpolating curve at reference points, at which the curve complies with the specified conditions. The characteristics of the curve are assigned within the area of their possible location. The possibilities of the proposed algorithm are investigated by interpolating the point series assigned to the branches of the parabola. In solving the test example, deviations of the normals and curvature radii from the corresponding characteristics of the original curve have been determined. The values obtained confirm the correctness of the solutions proposed in the paper.*

Keywords: interpolation, monotone curve, singular points, normal, centre of curvature, evolute, curvature radius

ULEPSZENIE ALGORYTMU WYZNACZANIE CHARAKTERYSTYKI INTERPOLACYJNEJ KRZYWEJ MONOTONICZNEJ

Streszczenie. *Interpolacja szeregu punktowego jest niezbędnym krokiem w rozwiązywaniu takich problemów, jak budowanie grafów opisujących zjawiska lub procesy, a także modelowanie w oparciu o zbiór punktów odniesienia układów liniowych definiujących powierzchnię. Aby uzyskać odpowiedni model, na interpolowaną krzywą stawia się następujące warunki: minimalną liczbę punktów osobliwych (punktów załamania, punktów przegięcia lub punktów skrajnej krzywizny) oraz regularną zmianę krzywizny wzdłuż krzywej. Celem pracy jest opracowanie algorytmu przypisania charakterystyk (położenia normalnych i wartości krzywizny) krzywej interpolacyjnej w punktach odniesienia, w których krzywa spełnia określone warunki. Charakterystyki krzywych nadawane są w obszarze ich możliwego położenia. Możliwości proponowanego algorytmu są badane poprzez interpolację szeregów punktów przypisanych do gałęzi paraboli. W rozwiązaniu przykładu testowego wyznaczono odchylenia normalnych i promieni krzywizny od odpowiednich charakterystyk pierwotnej krzywej. Otrzymane wartości potwierdzają poprawność zaproponowanych w pracy rozwiązań.*

Słowa kluczowe: interpolacja, krzywa monotoniczna, punkty osobliwe, normalna, środek krzywizny, ewolucja, promień krzywizny

Introduction

Interpolation of a series of fixed points is a necessary step in solving many geometric modelling problems. Such problems include creating graphs describing phenomena or processes, as well as modelling based on a set of reference points of the linear frameworks defining the surface.

The input data for process research are always discrete and can be presented on the graph by a sequence of points. Interpolation of reference points makes it possible to estimate the characteristics of the process at any point through the coordinates of the points of the interpolating curve. For this estimate to be correct, the interpolating curve configuration must comply with the layout of the sequence of reference points.

Such compliance is impossible without controlling the emergence of singular points along the interpolating curve. These can be:

- kinking points, where the curve has two tangent lines;
- inflection points, at which the convex and concave parts of the curve meet;
- extreme curvature points, where the direction of the increase in curvature along the curve changes;
- points at which the regularity of curvature changes along the curve is disturbed.

For a phenomenon or process to be adequately represented, the interpolation method must provide a number of singular points along the curve, minimum possible by task conditions. This means that the curve contains singular points in areas where, based on the configuration of the sequence of reference points, their presence is imperative. At the same time, there should be no singular points on sections that can be interpolated by a monotone curve along which curvature values change monotonously and regularly.

In case of uncontrolled emergence of singular points, the interpolating curve may deviate from the reference points to an uncontrolled distance. Therefore, the corresponding graph sections will not reflect the characteristics of the original phenomenon or process accurately.

Modelling complex surfaces is based on forming linear frameworks [5, 14, 25, 26]. In many cases, framework lines can only be obtained by interpolating a sequence of points. The number of reference points may be significant, and the sequence of points may have a complex setup. This happens when the input data for the surface model is an array of points whose coordinates are obtained by measuring an existing prototype (reverse engineering) [16, 24, 27, 28, 32]. Another example is grid-based modelling, where two families of lines intersect to define the array of points that must be interpolated.

Complex surfaces tend to bound products whose function is to interact with the environment. These are surfaces with elevated aero- or hydrodynamic properties. The laminar flow of such surfaces is ensured by the characteristics of the curves forming the surface framework [2, 6, 9, 10, 15, 18, 23]. The minimum by task conditions number of singular points of the curves forming the linear framework is the basic condition providing elevated dynamic properties of the surface.

Imposing additional conditions on the interpolating curve requires increasing the degree of its equation. Such conditions may include the number of points through which the curve passes, location of tangents, and curvature values of the curve at specific points. The more conditions are imposed on the curve, the higher the possibility that it will have singular points, and the harder it is to control their presence and location.

If the interpolating curve is formed as a polynomial [1], the degree of its equation is less by one than the number of reference points. If the polynomial degree is higher than 3,

it is impossible to control the emergence of inflection points. The emergence of inflection points can be prevented by interpolating second-order curves [3, 8, 19, 22, 30] or a B-spline [4, 17, 20, 21]. However, these methods are insufficient in terms of controlling the emergence of extreme curvature points.

The contradiction between imposing additional conditions on the interpolating curve and the control of the emergence of singular points on its sections is objective. The analytical representation of the curve lies at the basis of this contradiction. The equation defining the curve specifies its configuration, smoothness, and the pattern of changes in curvature values along the curve. If these patterns are out of compliance with the configuration of the reference points sequence and the characteristics assigned to them, then the emergence of singular points is inevitable. At present, this contradiction does not have a system solution.

Papers [11–13] propose a solution to the problem of controlling the emergence of singular points on the interpolating curve based on avoiding the analytical representation of its sections.

The author of [13] proposes a solution to the problem of forming a curve in the form of the area of possible location of its monotone parts. The area of possible location of the curve is formed as a sequence of closed contours, joined together at reference points. The dimensions of each of the contours are determined by the distance between the respective reference points, the absence of singular points on the section of the curve and the characteristics of the curve assigned at the reference points.

The area of possible location of section $i…i$+1 of the curve with no inflection points is defined by triangles i; N; i+1 (Fig. 1). The triangles are bounded by lines (i-1;i), (i;i+1) and (i+1;i+2) passing through three pairs of consecutive reference points.

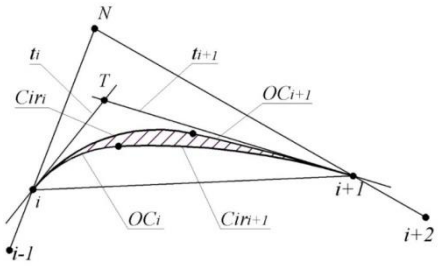


Fig. 1. Area of possible location of the interpolating curve

If the positions of the tangents are assigned to the reference points, the possible location of the interpolation curve section is triangle i; T; i+1, bounded by segment [i, i+1] and tangents t_i and t_i+1.

The area of possible location of the section of the curve with no extreme curvature points is bounded by arcs of circles. The boundaries of the area consist of arcs of circles osculating to the curve at reference points (OC_i and OC_{i+1}) and arcs of circles tangent to the osculating circle and a tangent at a different reference point (Cir_{i+1} and Cir_i) (Fig. 1).

Research has established that the value of the area of possible location of the curve, determined based on the assumption that there are no extreme curvature points, is 2-3 times less than the area determined by the convexity condition of the curve.

The analytical description of the boundaries of the area of possible location of the monotone curve was obtained in [11]. The article also offers a solution for the task of providing a given interpolation accuracy. The assignment of intermediate points within the reference sections corresponding to the tangent line and osculating circles leads to an increase in the number of sections and a decrease in their dimensions. The location area of the curve is considered to be formed when the dimensions of the maximum of its sections do not exceed the given value. In this case, the graph describing the phenomenon or process can be represented as an area of possible location of the curve.

While modelling a surface, lines contours are formed within the areas of the location of its framework, representing the curves with assigned characteristics with specified accuracy. The solution of the problem within the area of possible location of the monotone curve is proposed in [11]. The contour is formed by smoothly joined arcs of circles. Location of the contour within the specified area is provided by:

– the given direction of monotone increase of the radii of circles along the contour;
– the contour's contingency with the interpolating curve at reference points.

The disadvantage of solutions proposed in [11] is a complex algorithm for deter-mining the boundaries of the curve's location and a large number of necessary calculations.

The size of the areas of possible location of the interpolating curve is determined by the positions of its tangents and the curvature values assigned by the designer at reference points. The option of assigning these characteristics proposed in [13] ensures that there is such an area at each of the sections. The task of controlling the width ratio of the areas of different sections was not considered in the work. As a consequence, an area is likely to be formed where narrow sections alternate with areas of significant width. The curve formed within such an area will consist of areas, along which curvature changes rapidly, as well as areas where it changes insignificantly. Such a solution would reduce both the interpolation accuracy and the dynamic properties of the modelled surfaces.

The possibility to level the width of the adjacent sections of the area of the curve by correcting the characteristics of the interpolating curve at the point separating the sections was investigated in [11, 12]. The possibility to improve the obtained solution as a result of successive iterations was established.

Implementing the iterative approach in computer software implies the existence of consecutive cycles providing a step-by-step approximation of the curve's characteristics to the required values. The sequence of interpolated points can be in the thousands. Consequently, the iteration process can take a considerable amount of time, making it difficult to form the area of the curve's location in interactive mode.

The need for software implementation for the method under development re-quires increasing its effectiveness through:

1. developing a method for assigning the correct characteristics of the interpolating curve at reference points;
2. developing a simpler algorithm for determining the area of possible location of the monotone curve.

The article is aimed at the development and approbation of the algorithm for as-signing the positions of curvature centres corresponding to reference points, at which it is possible to provide regular and uniform change in curvature values along the interpolating curve. To achieve this aim, it is necessary to:

– develop a method of assigning the location of the normals based on the assumption that the curve has no singular points;
– develop a method for assigning the positions of the curvature centres on already assigned normals, which would provide a uniform change in curvature values along the curve;
– test the proposed algorithm when assigning curvature centres to interpolating curves at reference points assigned to a monotone curve.

The development of these methods requires solving the following tasks:

– determining the area of possible location of the normal of the interpolating monotone curve at reference points;
– determining the optimal location of each normal within the corresponding area, based on the conditions of the task;
– determining the areas of possible location of curvature centres and assigning their final positions.

1. Materials and methods

Let us consider how a sequence of curvature centres (C_i) of the monotone curve interpolating a given point series is formed, by the example of the curve with increasing radii of curvature.

The osculating circle corresponding to point i (OC_i) divides the monotone curve into two parts (Fig. 2).

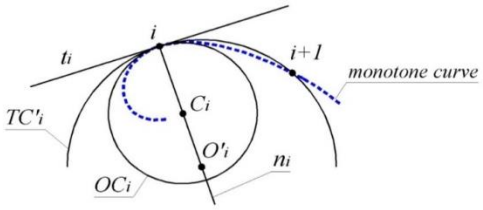


Fig. 2. Position of osculating and tangent circles with respect to the monotone curve

The part of the curve at which the radius of curvature is smaller than the radius of OC_i (R_i) is located inside it. The rest of the curve is outside OC_i. Point i, tangent to the curve at this point (t_i) and point $i+1$, located on the curve outside OC_i define the tangent circle TC'_i. The radius of TC'_i is greater than R_i. Point $i-1$ belonging to the curve and located inside OC_i defines $'TC_i$, whose radius is smaller than R_i. Points $i-1$, i, $i+1$ define the adjacent circle (AC_i) (Fig. 3). The radius of AC_i can be both larger and smaller than R_i.

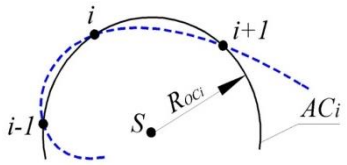


Fig. 3. Position of OC_i with respect to the monotone curve

The points assigned to the monotone curve define a sequence of adjacent, tangent, and osculating circles whose radii increase monotonously along the point series:

$$...<R\,AC_{i-1}<R\,TC'_{i-1}<R$$
$$TC_i<R\,AC_i<R\,TC'_i<R\;TC_{i+1}<R\,AC_{i+1}<... \quad (1)$$
$$...<R_{i-1}<R\,TC'_{i-1}<R\;'TC_i<R_i<R\,TC'_i<R\;'TC_{+1}<R_{i+1}<... \quad (2)$$

The position of the curvature centres of the interpolating monotone curve corresponding to the reference points shall be determined and assigned based on (1), (2).

The centres of AC … S_{i-1}, S_i, S_{i+1}, … are located at the intersection of the lines drawn through bisecting points of the segments (point K_i) perpendicularly to the segments connecting the adjacent reference points (Fig. 4).

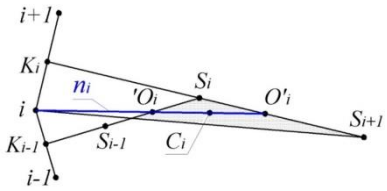


Fig. 4. Original area of the i-th centre of curvature

According to (1), the normal of the monotone curve at point i (n_i) must intersect sections $[S_{i-1}, S_i]$ and $[S_i, S_{i+1}]$ at points $'O_i$ and O'_i – centres of $'TC_i$ и TC'_i respectively. According to (2), points $'O_i$ and O'_i bound the range of the centre of curvature C_i on normal n_i. The original area of location of C_i is a triangle bounded by lines (K_{i-1}, S_i), (K_i, S_{i+1}) and that of lines (i, S_{i+1}) or (i, S_{i-1}), which is at a smaller angle to (K_{i-1}, S_i).

The reference criterion for the mutual arrangement of the normals assigned to adjacent points i and $i+1$ is the ratio of the lengths of the segments (Fig. 5):

$$|S_i, O'_i| \le |C_i, {}'O_{i+1}|. \quad (3)$$

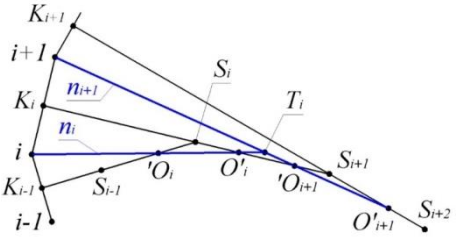


Fig. 5. Mutual arrangement of normals assigned to adjacent points

The equality of the specified segments means that point T_i belongs to line (K_i, S_i). In this case, section $(i, i+1)$ of the monotone curve is the arc of the circle with centre T_i.

When assigning the position of normals at reference points, we will strive for a uniform arrangement of the centres of TC within the segment bounded by the centres of the corresponding AC.

Let us look at the variant of assigning such positions to the normals.

1. Assign a preliminary position of n_i, at which point $'O_i \equiv n_i \times (K_{i-1}, S_i)$ separates segment $[S_{i-1}, S_i]$ at a ratio of $|'O_i, S_i| : |S_{i-1}, S_i| = 1{:}3$.
2. Determine point O'_i, at which n_i intersects line (K_i, S_i). If the value of the ratio of the lengths of segments $|S_i, O'_i| : |S'_i, S_{i+1}|$ does not exceed 1:2, then the position of n_i is considered to be finalized. Otherwise, we assign the final position of n_i, at which point O'_i divides segment $[S_i, S_{i+1}]$ at a ratio of $|O'_i, S_i| : |S_i, S_{i+1}| = 1{:}3$.
3. The position of normal n_{i+1} is selected from two options:

– point $'O_{i+1} \equiv n_{i+1} \times (K_i, S_{i+1})$ divides segment $[O'_i, S_{i+1}]$ into equal parts;
– point $O'_{i+1} \equiv n_{i+1} \times (K_{i+1}, S_{i+1})$ divides segment $[S_{i+1}, S_{i+2}]$ at a ratio of $|O'_{i+1}, S_{i+1}| : |S_{i+1}, S_{i+2}| = 1{:}3$.

The final position of n_{i+1} is that, at which the normal is closer to point S_{i+1}.

Having assigned the positions of normals at all reference points according to the given scheme, we obtain a sequence of centres of TC whose radii comply with condition (1). The criterion for the final selection of the positions of normal is the ratio of the lengths of the parts into which the normals divide the segments connecting the centres of the corresponding AC. The position of normals in which the specified segments are divided into three equal parts shall be considered optimal.

The position of the centres of curvature is determined on corresponding normals based on the properties of the evolute of the monotone curve [7, 29]:

– evolute is a convex curve with no inflection points or spinodes;
– the normals of the curve are tangent to its evolute in the respective centres of curvature;
– the length of any section of the evolute is equal to the difference in values of the radii of curvature at points bounding the corresponding section of the original curve.

The position of the centres of curvature C_i is assigned within the respective segments $['O_i, O'_i]$. Consider the option of assigning these positions.

1. For each section of the interpolating curve, the minimum length of the evolute is determined. For section $(i...i+1)$ such an evolute is segment $[C'_i, {}'C_{i+1}]$, the length and location of which are determined by the conditions:

$$\begin{cases} |C'_i, {}'C_{i+1}| = \Delta r_i; \\ |C'_i, T_i| = |{}'C_{i+1}, T_i|, \end{cases} \quad (4)$$

where $\Delta r_i = |i + 1, {}'C_{i+1}| - |i, + 1, C'_i|$ – the difference in the radii values of the circles whose arcs make up the section of contour $(i, i+1)$.

Segment $[C'_i, {}'C_{i+1}]$ can be defined by finding its position in relation to triangle $A, {}'O_{i+1}, T_i$ (Fig. 6).

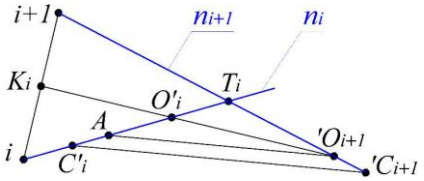


Fig. 6. Finding the minimum evolute of a section of the interpolating curve

Point A is assigned on line n_i based on the equality of segments $|A,T_i|=|'O_{i+1},T_i|$. The position of segment $[C'_i, 'C_{i+1}]$ in relation to triangle $A,'O_{i+1}, T_i$ determines the coefficient:

$$f = \frac{|C'_i,T_i|}{|A,T_i|} = \frac{|'C_{i+1},T_i|}{|'O_{i+1},T_i|} = \frac{|C'_i,'C_{i+1}|}{|A,'O_{i+1}|}, \quad (5)$$

whose value is calculated by the formula:

$$f = \frac{\Delta}{2|T_i,'O_{i+1}|-|A,'O_{i+1}|}+1, \quad (6)$$

where $\Delta=|i,A|+|A,'O_{i+1}|-|i+1,'O_{i+1}|$ is equal to the difference in lengths of evolute $[C'_i, 'C_{i+1}]$ and segment $[A,'O_{i+1}]$.

2. After determining the minimum evolutes for each section of the curve, a sequence of segments …, $[C'_{i-1},'C_i]$, $[C'_i,'C_{i+1}]$, … is obtained (Fig. 7) determining the involute as smoothly joined arcs of circles. This involute interpolates the reference point series by locating within the area of possible location of the monotone curve and is similar to the contour formed by method proposed in [11].

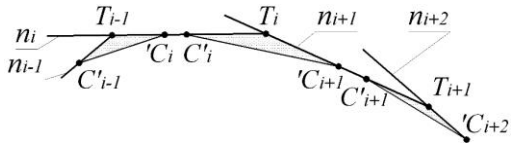


Fig. 7. Sequence of minimum evolutes

3. The curvature centres of the interpolating curve C_i, corresponding to the reference points, are assigned within the boundaries of segments $['C_i, C'_i]$. Assigning the i-th centre of curvature within the range of segment $[i, C'_i]$, and the $i+1$-th centre of curvature outside line $[i+1, 'C_{i+1}]$ enables forming the section of the evolute as a smooth convex line (Fig. 8), which defines the involute as a monotone curve with a regular curvature change.

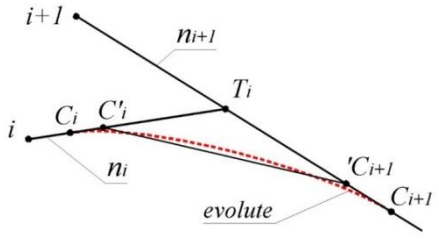


Fig. 8. Evolute of the section of the monotone curve

The criterion for the final assignment of curvature centres within the reference ranges $['O_i, O'_i]$ is the ratio of the lengths of the parts into which these centres divide the corresponding segments $['C_i, C_i']$. The position of the centres of curvature in the middle of these segments shall be considered optimal.

The normals of the interpolating curve assigned at reference points and the chords connecting the curvature centres assigned on these normals bound the sequence of triangles (Fig. 9).

The sides of each triangle correspond to:

$$|C_i, C_{i+1}| \le R_{i+1} - R_i \le |C_i, T_i| + |C_{i+1}, T_i|, \quad (7)$$

and their sequence is the area of possible location of the evolute of the monotone curve that interpolates the set reference points.

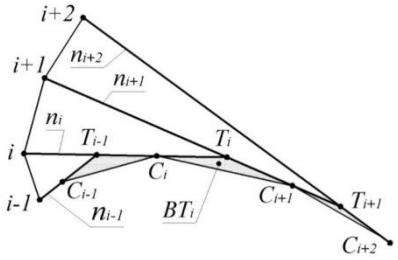


Fig. 9. Area of location of the evolute of the discretely presented curve

The solutions presented above allow assigning the characteristics of the interpolating curve at reference points according to the following algorithm:

1. The original point series is divided into parts that can be interpolated by a mono-tone curve.
2. For the reference points defining the monotone sections of the interpolating curve, the areas of location of normal are determined, at which it is possible to ensure the absence of singular points for these sections.
3. The position of each of the normals is determined based on the greatest possible approximation of the criterion determining its location within the area to the value accepted as optimal.
4. The ranges of possible location of the centres of curvature are determined on the assigned normals, based on the task of ensuring a uniform change of curvature values along the curve.
5. The centres of curvature are assigned within the respective ranges, based on the given ratio of lengths of the segments into which the range is divided by the centre of curvature.

2. Results and discussion

Consider solving the problem of assigning curvature centres to an interpolating curve based on forming the area of possible location of the evolute of the monotone curve, on the example of a sequence of reference points assigned to the branch of the parabola defined by the equation $y = \frac{x^2}{300}$. The characteristics of the reference point series are given in table 1.

Table 1. Characteristics of the reference point series

i	Point coordinates, mm		Chord length $\|i, i+1\|$, mm	Radius of AC RAC_i, mm	Range of location of the normal Δn_i, °
	x	y			
1	30	3	31.32	-	1.046
2	60	12	33.54	188.58	1.074
3	90	27	36.62	238.73	1.074
4	120	48	84.85	348.37	1.386
5	180	108	103.23	570.09	2.727
6	240	192	123.55	1002.64	2.244
7	300	300	-	-	-

The increasing values of the radii of AC along the sequence of reference points determine the direction in which the radii of curvature increase along the monotone curve, which can interpolate these points.

The values given in the sixth column of the table of the ranges of the normals of the interpolation curve are determined by the smaller of the angles between line (i, S_i) and one of the lines (i, S_{i-1}) or (i, S_{i+1}) (Fig. 4). For example, for point 4, range Δn_4 is equal to angle $S_4, 4, S_5$. For point 1, the range is defined by angle $B, 1, S_2$, where B is the intersection point of $(2, S_3)$ and (K_1, S_2). For the last point, the normal range is not determined because the AC_7 does not exist, and point S_7 does not bound the turn of normal n_7 towards increasing angle $6, 7, 'O_7$. The equality of ranges Δn_2 and Δn_3 means that their value is determined by the condition of intersection by normals n_2 and n_3 of segment $[S_2, S_3]$.

The position assigned to each of the normals determines the ratio of the lengths of the segments:

$$K_i = \frac{|S_i; O_i'|}{|S_i; S_{i+1}|}, \quad (8)$$

where O'_i – the intersection point of normal n_i with line (K_i,S_i) (Fig. 5).

The coefficient values of K_i for the normals of the interpolating curve are given in Tab. 2.

Table 2. Characteristics of the reference point series

i	1	2	3	4	5	6	7
K_i	-	0.333	0.292	0.283	0.301	0.276	-

The position of normal n_2, which corresponds to the smallest initial range of possible location, is assigned first. The position of normals n_3… n_6 is defined according to the above method. The position of the normal n_1 at the first reference point is determined by the ratio of the lengths of segments $|O'_1, {}'O_2| = |{}'O_2, S_2|$. The position of normal n_7 at the last, seventh point is determined by the ratio of the lengths of segments $|S_6, O'_6| = |O'_6, {}'O_7|$.

Similar coefficients values of K_i reflect the correct assignment of the position of normals, which enables a monotonous and uniform change in the curvature values along the formed curve.

Once the normals have been assigned at reference points, the dimensions and position of the minimum evolutes are determined for each section of the interpolating curve, and then the curvature centres corresponding to the reference points are assigned. The results are shown in Tab. 3.

The proposed method makes it possible to determine the boundaries of the ranges of location of the curvature centre for all the reference points except the first and last ones. For the first point, the upper boundary of ($'C_1$), the extreme position of which is reference point 1 itself, is not determined. For the last point, the lower boundary of (C'_7), which can be at an arbitrary large distance from reference point 7, is not determined. We shall define the radius of the curvature centre based on the ratio of the lengths of similar ranges to the adjacent sections, as follows:

$$|{}'C_1, C'_1| = |{}'C_2, C'_2| \frac{|{}'C_2, C'_2|}{|{}'C_3, C'_3|}, \quad (9)$$

$$|{}'C_7, C'_7| = |{}'C_6, C'_6| \frac{|{}'C_6, C'_6|}{|{}'C_5, C'_5|}. \quad (10)$$

The parabola, on which a sequence of reference points was assigned, can be considered as a variant of the monotone curve that can interpolate these points. In the seventh column of the table, for comparison, the curvature radii of the parabola at reference points are given.

Deviation of the positions of normals and the relative deviation of the curvature radii for the interpolating curve are shown in Tab. 4.

The greatest deviation from the characteristics of the original curve occurred at the first and the last reference points. The reason for this error may be the fact that one of the boundaries of the characteristic ranges corresponding to these points is calculated based on the coordinates of the rest of reference points, and the other boundary is assigned intuitively, based on logical reasoning. This error can be reduced by reducing the distances between the reference points at the beginning and at the end of their sequence by increasing their number [11, 31].

The similarity of the values of the characteristics assigned by the proposed meth-od to the corresponding characteristics of the original curve confirms the correctness of the results presented in the article.

Table 3. Characteristics of the reference point series

i	Radius of TC, mm		Minimum evolute length, mm	Centre of curvature range, mm	Curvature radius of the interpolating curve	Curvature radius of the parabola R_i^{par}, mm
	$R'TC_i$	$R\ TC'_i$	$\|C'_b{}'C_{i+1}\|$	$\|{}'C_b C'_i\|$	R_i, mm	
1	-	166.74	29.46	5.30	155.10	159.09
2	176.10	201.94	55.23	10.02	184.56	187.40
3	218.66	260.65	110.82	18.94	239.79	237.90
4	319.22	384.88	218.92	40.58	350.61	355.03
5	468.62	693.55	417.33	75.50	569.53	571.71
6	837.83	1200.0	742.18	134.33	986.86	1007.55
7	1422.14	-	-	239.00	1729.04	1677.05

Table 4. Deviation of characteristics of the interpolating and the reference curves

Point number, i	1	2	3	4	5	6	7
Deviation angle of the normal, °	0.32	0.14	0.11	0.23	0.18	0.25	1.07
Deviation of the curvature radius, %	2.51	1.50	0.79	1.24	0.38	2.05	3.10

3. Conclusions

The article offers an algorithm for assigning the positions of curvature centres of the interpolating curve at reference points. The curvature centres are assigned on the basis of a regular and uniform increase in curvature values along the sections of the curve where the initial conditions prevent the occurrence of singular points.

The algorithm is based on the following methods:

- assigning the positions of normals of the interpolating curve at reference points;
- assigning the positions of curvature centres on the assigned normals.

The position of each of the normals is assigned within an area, the boundaries of which are defined by the specified properties of the interpolating curve. The reference area of the possible location of the normal is uniquely defined by the coordinates of five consecutive reference points. The boundaries of the reference area shall be refined in accordance with the condition that the normals are assigned simultaneously at the previous and subsequent reference points. The proposed method for simultaneous assignment of normals to all the reference points provides:

- a proportional decrease of the reference area of the normal for all the points;
- location of each normal in the centre of the refined area.

These proportions are the compliance criterion of the assigned normals with the configuration of the reference point series.

The position of each centre of curvature is assigned within a predetermined segment that belongs to the corresponding normal. These segments are the ranges of curvature centres, and they take into account the whole area of a possible solution with the assigned positions of the normals. The proportions in which the curvature centres divide the corresponding ranges may serve as a criterion for the correctness of assigning their positions.

Assigning the centres of curvature at the points bounding the mentioned ranges provides a unique solution – an interpolating curve consisting of smoothly joined arcs of circles whose radii increase monotonously along the curve. Assigning the curvature centres within the ranges determines the possible location of the evolutes of the mono-tone regular curve that interpolates

the reference points. The position of the centres of curvature and the points bounding the ranges of their possible assignment uniquely determine the possible location of the involute – the interpolating curve.

In this case, the solution is not unique. Based on the same location of the involute, it is possible to form a set of interpolating curves, whose characteristics comply with the conditions of the problem.

The possibilities of the proposed algorithm and its constituent methods have been investigated by interpolating a sequence of points assigned to the branches of the parabola. The standard positions of normals obtained by solving the test example deviate from the corresponding normals of the reference curve (parabola) within 0.32 degrees. The relative deviation of the assigned curvature radii from the corresponding values of the reference curve was within the range of 0.79–3.10 per cent. The values of these deviations confirm the correctness of the proposed solutions.

The proposed methods are based on geometric constructions, which resolve themselves to determining the intersection points of straight lines and dividing the segments in the fixed ratio. The necessary calculations consist in solving systems of linear equations. The simplicity of the geometric and computational schemes ensures high accuracy of calculations and does not require the application of iteration pro-cesses.

These features make the proposed algorithm the most appropriate for the task of its further implementation in the form of a computer program.

The disadvantages of the proposed solutions include the fact that the calculations of the division ratio of the initial ranges are based on logical reasoning and are currently insufficiently investigated. These ratios provided a positive result in solving the test example, but they cannot be established as optimum. The elimination of this deficiency requires further investigation of various sequences of points assigned on various curves.

The results obtained in this work complement and develop the research carried out in previous works [11–13]. Assigning the characteristics of the interpolation curve based on the area of possible location of its evolute simplifies the interpolation problem, reduces computational error, making the solution more reliable. However, this paper does not address the issue of ensuring the given accuracy of interpolation. The problem of forming the evolute of the curve, which contains a minimum based on the initial data number of singular points and interpolates a sequence of reference point with given accuracy, is to be solved in further research.

References

[1] Argyros I. K., George S.: On the convergence of Newton-like methods restricted domains. Numer. Algorithms 75(3), 2017, 553-567 [http://doi.org/10.1007/s11075-016-0211-y].

[2] Bucsa S., Serban A., Balan M. C., Ionita C., Nastase G., Dobre C., Dobrovicescu A.: Exergetic Analysis of a Cryogenic Air Separation Unit. Entropy 24, 2022, 272 [http://doi.org/10.3390/e24020272].

[3] Chekalin A. A., Reshetnikov M. K., Shpilev V. V., Borodulina S. V.: Design of Engineering Surfaces Using Quartic Parabolas. IOP Conf. Ser.: Mater. Sci. Eng. 2007, 012015 [http://doi.org/10.1088/1755-1315/221/1/012015].

[4] Farin G., Rein G., Sapidis N., Worsey A. J.: Fairing cubic B-spline curves. Computer Aided Geom. Des. 4(1–2), 1987, 91-103 [http://doi.org/10.1016/0167-8396(87)90027-6].

[5] Fooladi M., Foroud A. A.: Recognition and assessment of different factors which affect flicker in wind turbine. IET Renew. Power Gener. 1, 2015, 250–259 [http://doi.org/10.1049/iet-rpg.2014.0419].

[6] Halko S., Halko K., Suprun O., Qawaqzeh M., Miroshnyk O.: Mathematical Modelling of Cogeneration Photoelectric Module Parameters for Hybrid Solar Charging Power Stations of Electric Vehicles. IEEE 3rd KhPI Week on Advanced Technology, Kharkiv, 2022, 1-6 [http://doi.org/10.1109/KhPIWeek57572.2022.9916397].

[7] Halko S., Suprun O., Miroshnyk O.: Influence of temperature on energy performance indicators of hybrid solar panels using cylindrical cogeneration photovoltaic modules. IEEE 2nd KhPI Week on Advanced Technology, 2021, 21259624, 132–136 [http://doi.org/10.1109/KhPIWeek53812.2021.9569975].

[8] Hashemian A., Hosseini S. F.: An integrated fitting and fairing approach for object reconstruction using smooth NURBS curves and surfaces. Comput. Math. with Appl. 76(7), 2018, 1555-1575 [http://doi.org/10.1016/j.camwa.2018.07.007].

[9] Hashemian A., Imani B. M.: Surface fairness: a quality metric for aesthetic assessment of compliant automotive bodies. J. Eng. Des. 29(1-2), 2018, 41-64 [http://doi.org/10.1080/09544828.2018.1435853].

[10] Havrylenko Y., Cortez J. I., Kholodniak Y., Alieksieieva H., Garcia G. T.: Modelling of surfaces of engineering products on the basis of array of points. Teh. Vjesn. 27(6), 2020, 2034–2043 [http://doi.org/10.17559/TV-20190720081227].

[11] Havrylenko Y., Kholodniak Y., Halko S., Vershkov O., Bondarenko L., Suprun O., Miroshnyk O., Shchur T., Śrutek M., Gackowska M.: Interpolation with Specified Error of a Point Series Belonging to a Monotone Curve. Entropy 23(5), 2021, 493 [http://doi.org/10.3390/e23050493].

[12] Havrylenko Y., Kholodniak Y., Halko S., Vershkov O., Miroshnyk O., Suprun O., Dereza O., Shchur T., Śrutek M.: Representation of a Monotone Curve by a Contour with Regular Change in Curvature. Entropy 23(7), 2021, 923 [http://doi.org/10.3390/e23070923].

[13] Havrylenko Y., Kholodniak Y., Vershkov O., Naidysh A.: Development of the method for the formation of one-dimensional contours by the assigned interpolation accuracy. East.-Eur. J. Enterp. Technol. 1(4(91)), 2018, 76-82 [http://doi.org/10.15587/1729-4061.2018.123921].

[14] Hosseini S. F., Moetakef-Imani B.: Innovative approach to computer-aided design of horizontal axis wind turbine blades. J. Comput. Des. Eng. 4(2), 2017, 98-105 [http://doi.org/10.1016/j.jcde.2016.11.001].

[15] Karaiev O., Bondarenko L., Halko S., Miroshnyk O., Vershkov O., Karaieva T., Shshur T., Findura P., Pristavka M.: Mathematical modelling of the fruit-stone culture seeds calibration process using flat sieves. Acta Technologica Agriculturae 24(3), 2021, 119–123 [http://doi.org/10.2478/ata-2021-0020].

[16] Ke Y., Fan S., Zhu W., Li A., Liu F., Shi X.: Feature-based reverse modeling strategies. Comput. Aided Des. 38(5), 2006, 485-506.

[17] Lan P., Yu Z., Du L., Lu N.: Integration of non-uniform Rational B-splines geometry and rational absolute nodal coordinates formulation finite element analysis. Acta Mech. Solida Sin. 27(5), 2014, 486-495 [http://doi.org/10.1016/S0894-9166(14)60057-4].

[18] Lee T.-W., Park J. E.: Entropy and Turbulence Structure. Entropy 24, 2022, 11 [http://doi.org/10.3390/e24010011].

[19] Li H.: Geometric error control in the parabola-blending linear interpolator. J. Syst. Sci. Complex. 26(5), 2013, 777-798 [http://doi.org/10.1007/s11424-013-3178-y].

[20] Li W., Xu S., Zheng J., Zhao G.: Target curvature driven fairing algorithm for planar cubic B-spline curves. Computer Aided Geom. Des. 21(5), 2004, 499-513 [http://doi.org/10.1016/j.cagd.2004.03.004].

[21] Okaniwa Sh., Nasri A., Lin H., Abbas A., Kineri Yu., Maekawa T.: Uniform B-Spline Curve Interpolation with Prescribed Tangent and Curvature Vectors. IEEE Trans. Vis. Comput. Graph. 18(9), 2016, 1474-1487 [http://doi.org/10.1109/TVCG.2011.262].

[22] Park H., Kim K., Lee S-C.: A method for approximate NURBS curve compatibility based on multiple curve refitting. Comput. Aided Des. 32(4), 2000, 237-252 [http://doi.org/10.1016/S0010-4485(99)00088-3].

[23] Pazyi V., Miroshnyk O., Moroz O., Trunova I., Savchenko O., Halko S.: Analysis of technical condition diagnostics problems and monitoring of distribution electrical network modes from smart grid platform position. IEEE KhPI Week on Advanced Technology, Kharkiv, 2020, 57-60 [http://doi.org/10.1109/KhPIWeek51551.2020.9250080].

[24] Peng Y. H., Yin Z. W.: The algorithms for trimmed surfaces construction and tool path generation in reverse engineering. Comput. Ind. Eng. 54(3), 2008, 624-633 [http://doi.org/10.1016/j.cie.2007.09.012].

[25] Pérez-Arribas F., Pérez-Fernández, R.: A B-spline design model for propeller blades. Adv. Eng. Softw. 118, 2018, 35–44 [http://doi.org/10.1016/j.advengsoft.2018.01.005].

[26] Pérez-Arribas F., Trejo-Vargas I.: Computer-aided design of horizontal axis turbine blades. Renew. Energ. 44, 2012, 252-260 [http://doi.org/10.1016/j.renene.2012.01.100].

[27] Piegl L. A., Tiller W.: Reducing control points in surface interpolation. IEEE Comput. Graph. Appl. 20(5), 2000, 6698012, 70-75 [http://doi.org/10.1109/38.865883].

[28] Qawaqzeh M., Szafraniec A., Halko S., Miroshnyk O., Zharkov A.: Modelling of a household electricity supply system based on a wind power plant. Przegląd Elektrotechniczny 96, 2020 [http://doi.org/10.15199/48.2020.11.08].

[29] Robbin J. W., Salomon D. A.: Introduction to Differential Geometry. Springer Spektrum, Zürich 2022.

[30] Shen W., Wang G., Huang F.: Direction monotonicity of a rational Bézier curve. Appl. Math. J. Chin. Univ. 31(1), 2016, 1–20 [http://doi.org/10.1007/s11766-016-3399-7].

[31] Szafraniec A., Halko S., Miroshnyk O., Figura R., Zharkov A., Vershkov O.: Magnetic field parameters mathematical modelling of windelectric heater. Przeglad elektrotechniczny 97(8), 2021, 36-41 [http://doi.org/10.15199/48.2021.08.07].

[32] Tabor S., Lezhenkin O., Halko S., Miroshnyk O., Kovalyshyn S., Vershkov O., Hryhorenko O.: Mathematical simulation of separating work tool technological process. 22nd International Scientific Conference on Progress of Mechanical Engineering Supported by Information Technology – POLSITA 2019, Czajowice, 2019, 132 [http://doi.org/10.1051/e3sconf/201913201025].

Ph.D. Eng. Yuliia Kholodniak
e-mail: yuliya.kholodnyak@tsatu.edu.ua

Candidate of Technical Sciences Yuliia Kholodniak, graduated from Tavria State Agrotechnological University, Ukraine, in 2011, qualified as Computer Systems Analyst. In 2016, she received a Ph.D. in technical sciences (applied geometry, engineering graphics) at Bogdan Khmelnitsky Melitopol State Pedagogical University, Ukraine. Now she is associate professor at Dmytro Motornyi Tavria State Agrotechnological University and head of the Department of Computer Science, Melitopol, Ukraine. Author of more than 100 scientific publications. Her research interests are related to the field of computer modelling of curved lines with given properties.

http://orcid.org/0000-0001-8966-9269

D.Sc. Eng. Yevhen Havrylenko
e-mail: yevhen.havrylenko@tsatu.edu.ua

Doctor of Technical Sciences Yevhen Havrylenko, graduated from Melitopol Institute of Agricultural Mechanization in 1984 with a degree in Agricultural Mechanization. Qualification: Mechanical Engineer of Agriculture. In 2004, he received Ph.D. in technical sciences (applied geometry, engineering graphics), and in 2020 received a doctorate in technical sciences at Kyiv National University of Construction and Architecture, Ukraine. Now he is Professor at Dmytro Motornyi Tavria State Agrotechnological University, Melitopol, Ukraine. He is the author of more than 180 scientific publications. His research interests are related to research in the field of computer modelling of curved lines with given properties.

http://orcid.org/0000-0003-4501-445X

Ph.D Eng. Serhii Halko
e-mail: serhii.halko@tsatu.edu.ua

Graduated from Melitopol Institute of Mechanical Engineers of Agriculture, Ukraine, in 1994, and was qualified as an Electrical Engineer. He received his Ph.D. in Electric Engineering (power stations, systems and networks) from the Donetsk National Technical University, Ukraine, in 2003. Presently, he is an assistant professor of Dmytro Motornyi Tavria State Agrotechnological University and Dean of the Faculty of Energy and Computer Technologies, Melitopol, Ukraine. He is the author of more than 100 scientific publications, 10 inventions, two monographs. His research interests are related to research in the field of renewable and alternative energy.

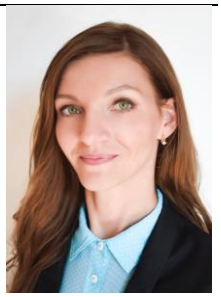
http://orcid.org/0000-0001-7991-0311

D.Sc. Eng. Volodymyr Hnatushenko
e-mail: vvgnat@ukr.net

Doctor of Technical Sciences Volodymyr Hnatushenko graduated from Dnepropetrovsk State University in 1999 with a degree in Technology and Telecommunications. In 2003 he received the degree of Candidate of Technical Sciences, and in 2009 he defended his thesis for the degree of Doctor of Technical Sciences in the speciality "Applied geometry, Engineering graphics". Now he is Professor at the National Technical University "Dnipro Polytechnic" and head of the Department of Information Technology and computer engineering, Dnipro, Ukraine. He is the author of more than 350 scientific and educational works, including patents and copyright certificates for inventions. He participated in numerous international conferences and symposia. His research interests are related to the field of information technology, mathematical and computer modelling of processes and systems.

http://orcid.org/0000-0003-3140-3788

Ph.D Eng. Olena Suprun
e-mail: olena.suprun@tsatu.edu.ua



Graduated from Melitopol State Teacher-Training University, Melitopol, Ukraine, in 2003 with a degree in Pedagogy and Methods of Teaching Languages and Literature (English, German) and was qualified as Teacher of English, German and Foreign Literature. Presently, she is a Senior Teacher of the Department of Foreign Languages at Dmytro Motornyi Tavria State Agrotechnological University, member of TESOL. She is the author of over 20 scientific publications and 2 books, 6 copyright certificates. Her research interests include innovative methods of teaching foreign languages as well as linguistic peculiarities of technical and engineering texts.

http://orcid.org/0000-0003-4369-712X

Ph.D Eng. Tatiana Volina
e-mail: t.n.zaharova@ukr.net

Candidate of Technical Sciences Tatiana Volina, she graduated from the Engineering and Technology Department of Sumy National Agrarian University in 2010 with a degree in Agricultural Mechanization. In 2014 she defended her thesis (specialty "Applied geometry, Engineering graphics"). Now she is an Associate Professor of the Department of Descriptive Geometry, Computer Graphics and Design of the National University of Life and Environmental Sciences of Ukraine, Kyiv. She is the author of about 115 scientific papers, 9 patents of Ukraine for a utility model. Her research interests are related to the field of geometric-kinematic methods for determining the parameters of particle motion on surfaces.

http://orcid.org/0000-0001-8610-2208

D.Sc. Eng. Oleksandr Miroshnyk
e-mail: omiroshnyk@btu.kharkiv.ua

Graduated from Kharkiv State Technical University of Agriculture, Kharkiv, Ukraine, in 2004, and was qualified as an Electrical Engineer. He received his Ph.D. in Electric Engineering (power stations, systems and networks) from the National Technical University "Kharkiv Politechnic Institute", Ukraine, in 2009 and Doctor of Technical Sciences degree in 2016. His research interests are related to research in the field of smart grid, renewable energy and computer numerical control machines.

http://orcid.org/0000-0002-6144-7573

Ph.D. Eng. Taras Shchur
e-mail: shchurtg@gmail.com

Graduated from Kharkiv State Technical University of Agriculture, Kharkiv, Ukraine, in 2004, and was qualified as a Mechanical Engineer. He received his Ph.D. in Mechanical Engineering (Machines and means of mechanization of agricultural production) from the Kharkiv State Technical University of Agriculture, in 2009. His research interests are related to research in the field of smart grid, renewable energy and computer numerical control machines.

http://orcid.org/0000-0003-0205-032X

http://doi.org/10.35784/iapgos.4459 *received: 30.08.2023 | revised: 09.10.2023 | accepted: 01.12.2023 | available online: 20.12.2023*

AN ANALYSIS OF THE IMPLEMENTATION OF ACCESSIBILITY TOOLS ON WEBSITES

Marcin Cieśla, Mariusz Dzieńkowski
Lublin University of Technology, Department of Computer Science, Lublin, Poland

Abstract. *The websites of higher education institutions, due to the fact that they are addressed to multiple stakeholder groups, not only need to have an appropriately designed information structure but must also be useful. Additionally, in the case of public universities, their services are expected to be accessible to the widest possible audience, especially for people with disabilities. The accessibility tools used on websites should be quickly located, easily identifiable and user-friendly. So far, no standards have been developed regarding these issues, and therefore, there are various solutions on the web. The objective of this study is to analyze various implementations of accessibility tools on university websites in terms of their location, form of presentation and ways that enable access to them. A study was conducted in which web interfaces were evaluated with the participation of users. The experiment consisted of two parts: the first one used the eye tracking technique, whereas in the second one, a survey was conducted. The research material was prototypes of websites from four different universities. Each website had two versions differing in implementation of accessibility tools. In the study, 35 participants were divided into two groups of people. Each group was shown one of the two sets of website prototypes and the users were tasked with finding and activating a specific accessibility tool. After exploring the websites, each participant completed a questionnaire that pertained to their opinions regarding aspects such as appearance, placement and a way to access tools dedicated to people with disabilities. The obtained data, processed to the form of heatmaps and fixation maps, were subjected to a qualitative analysis. The survey results and eye tracking data were analyzed quantitatively. On the basis of performed analyzes it can be concluded that the following factors have an impact on the reduction in efficiency and productivity of users: placement of accessibility tools on university websites in a place other than the upper right corner, an indirect access to these tools or their non-standard appearance.*

Keywords: web accessibility, web usability, eye tracking, A/B tests

ANALIZA IMPLEMENTACJI NARZĘDZI DOSTĘPNOŚCI NA STRONACH WWW

Streszczenie. *Strony uczelni wyższych ze względu na fakt, że są skierowane do wielu grup interesariuszy, oprócz tego, że muszą mieć odpowiednio zaprojektowaną strukturę informacji, to muszą być także użyteczne. W przypadku publicznych uczelni wyższych oczekuje się, że ich serwisy będą dostępne dla jak największego grona odbiorców, w szczególności dla osób z niepełnosprawnościami. Stosowane w serwisach narzędzia dostępności powinny być szybko lokalizowane, łatwo identyfikowane i proste w użyciu. Jak dotąd nie opracowano standardów dotyczących tych kwestii i w związku z tym istnieje w sieci wiele rozwiązań. Celem pracy jest analiza różnych implementacji narzędzi dostępności w serwisach WWW szkół wyższych pod względem ich rozmieszczenia, formy prezentacji oraz sposobów, które umożliwiają do nich dostęp. Zrealizowano badania, w których dokonano oceny interfejsów webowych z udziałem użytkowników. Eksperyment składał się z dwóch części: w pierwszej wykorzystano technikę eyetrackingową, natomiast w drugiej ankietowanie. Materiał badawczy stanowiły prototypy stron internetowych czterech różnych uczelni wyższych. Każdą ze stron opracowano w dwóch wersjach różniących się implementacją narzędzi dostępności. W badaniu wzięło udział 35 osób podzielonych na dwie grupy i zastosowano testy A/B. Każdej grupie zaprezentowano jeden z dwóch zestawów prototypów stron, a użytkownicy musieli odnaleźć i wykorzystać konkretne narzędzia dostępności. Następnie badani wypełniali ankietę, która dotyczyła preferencji odnośnie sposobu prezentacji narzędzi dedykowanych osobom niepełnosprawnym, w tym aspektów dotyczących wyglądu, rozmieszczenia oraz sposobu dostępu do nich. Pozyskane dane przetworzone do postaci map cieplnych i map fiksacji poddano analizie jakościowej. Wyniki z ankiet oraz dane eyetrackingowe przeanalizowano w sposób ilościowy. Na podstawie przeprowadzonych analiz możliwe jest stwierdzenie, że następujące czynniki mają wpływ na spadek efektywności oraz produktywności użytkowników: rozmieszczenie narzędzi dostępności na stronach uczelni w miejscu innym niż prawy górny róg, pośredni dostęp do narzędzi, czy też niestandardowy wygląd.*

Słowa kluczowe: dostępność stron internetowych, użyteczność stron internetowych, eyetracking, testy A/B

Introduction

Accessibility of web applications is one of the frequently discussed topics in recent years. With the increasing number of people using online services, there is also a growing population of users with disabilities. An important aspect of application development is considering their implementation in terms of adapting functionality to the needs of users, particularly those who may have difficulties accessing information.

With the development and popularization of internet services, there has been increased awareness among programmers and interface designers about the need to provide solutions, that facilitate user access to information. A breakthrough step was the definition of Web Content Accessibility Guidelines (WCAG) [2] issued by the Web Accessibility Initiative in 1999. They are based on four principles: perceivability, functionality, understandability and robustness, which describe the goals and practices of presenting content on websites in an accessible manner to users.

Web developers implement solutions that make it easier for users to access website resources. Among these are elements referred to as accessibility tools. Thanks to their use, the accessibility of information is continuously expanded to an increasingly wider audience, including individuals with disabilities. In order to unify and ensure equal access do data, many documents were defined specifying the requirement to use tools, including to change font type, contrast, displaying tooltips and text-to-speech capabilities. In Poland, one of such documents is the applicable Act from 2019 on the accessibility of applications of public entities, which must comply with the WCAG 2.1 guidelines at level AA.

Websites of state universities belong to the group of services that are required to ensure accessibility. Most universities follow their own practices when creating applications, such as different designs or arrangement of accessibility tools. In addition to ensuring functionality, a crucial aspect of interface design is implementing it in a clear and understandable manner for the users. A common challenge for users is the inconsistent appearance of buttons that perform the same functions. The diverse approaches of web developers to ensure usability stem from the lack of clearly defined guidelines for designing user-friendly websites.

The usability of the interface can be achieved through the use of developed design methods. An approach gaining popularity is user-centric design described by Donald Norman and defined in 1999 as ISO 13407:1999. The software development process is based on the study of a group of people, who are end users, in terms of requirements, assessment and experience related to their experiences with the system, application or service. In the case of usability assessment, the most important of the four stages of the design process is the last one, namely testing. It is based on the collection and verification of the respondents' experiences of using the application according to the developed tasks.

The aim of the work was to analyze public websites of universities in terms of the impact of the implementation of accessibility tools on the effectiveness of user interaction. At the beginning, a review of existing solutions on 20 selected

university websites was carried out and the most commonly used approaches for presenting accessibility tools were identified. Afterwards, four page templates were prepared in two versions of the implementation of accessibility tools with regard to the arrangement, colors and applied graphic symbols of icons. A research scenario consisting of tasks concerning two versions of pages with different implementation of accessibility elements was developed. The scenario was used during the eye tracking study. Following that, participants of the experiment completed a questionnaire pertaining to their preferences regarding the presentation and implementation of accessibility tools. Based on the analyses, it turns out that a large number of websites have accessibility problems and do not meet the guidelines [13].

1. Literature review

Numerous studies have been conducted selected websites were evaluated according to defined accessibility recommendations. The authors of the paper [1] presented an empirical analysis of five selected government websites. A group of 25 visually impaired individuals was involved in the study. Each participant was tasked with completing the prepared instructions covering the issues of accessibility and assessing the difficulty of their performing on the basis of a five-point scale.

An important and growing group of people are elderly users. Some adults struggle with difficulties arising from, among other things, deterioration of vision, motor skills, and cognitive abilities [7]. Based on conducted studies [2, 9] regarding activities on educational websites and the activity of Internet creators, the need to provide solutions for easy navigation on the web has been justified. In addition to the issue of accessibility and the issue of usability is also emphasized. It mainly concerns the way content is presented and the implementation of functionality in such a way that it facilitates the use, navigation and retrieval of information by users [8]. The paper [6] discusses aspects of web application development based on a user-centered approach. The layout and content organization of the page template were analyzed in terms of navigation and content arrangement. The importance of two issues was emphasized: the consistency of the arrangement of elements and their intuitiveness. The significance of implementing data search functionality was also highlighted.

There are different views among researchers regarding the issues of accessibility and usability. According to some, accessibility derives from usability, as shown in the work [12]. In addition to defining and analyzing usability, a large portion of research focuses on methods of evaluating it. The most commonly used approaches are heuristic analysis and user testing [4]. Both approaches complement each other and allow to detect different problems. Unlike heuristics, checklists allow detection of accessibility errors such as visibility of elements or memorability through interaction with the application. The paper [5] presented the results of a review on the usage of evaluation methods and it was found that in as many as 69% of cases, manual methods such as surveys were employed.

One of the increasingly used methods of usability evaluation is the eye tracking technique. It involves recording eye movement and collecting data on how users focus their eyes on elements of the website's or application's content. To conduct the study, a research group is engaged with the task of carrying out designated instructions involving various activities, such as finding selected information [3, 11]. The results of the experiment are presented in the form of heat maps, which makes it easy to determine the most visited places within the site and verify the actions taken by the user. Unfortunately, there are also doubts about the accuracy of the data obtained [10]. The recorded activity does not always imply intentional behaviour, but only drawing attention to a particular functionality or interesting design.

Both accessibility and usability are increasingly discussed topics in the field of web applications. Currently, many interfaces of websites and web applications, as well as mobile applications, are analyzed for usability and compliance with the developed guidelines. Various methods are used for this purpose, often combining two methods to increase the credibility of the evaluation.

2. Methods

The usability study of accessibility tools was divided into three stages. The first one of them was the analysis of 20 selected websites of higher education institutions in terms of the diversity of presentation of those tools. Based on the identified approaches, interface prototypes were developed, taking into account the applied solutions. Then, tasks were prepared to be performed by users during the designed experiment using the eye tracking technique. In the last part, the respondents filled out a questionnaire, which concerned the evaluation of the implemented solutions and the ease of interaction with the prototypes of the pages into which these tools were embedded.

2.1. Review of existing solutions

The main criteria considered in the review of the applied solutions were the layout, color scheme and the way they were presented in the form of icons identifying the various tools and their functions. Some of the most common solutions were implementations of tools that allow users to change font size, contrast and content search. Less common ones included link highlighting, modifying word spacing, friendly text, reading guide or sign language interpreter (table 1).

Table 1. The number of university websites that include individual accessibility tools varies

Accessibility tool	The number of implementations
Contrast	18
Font size adjustment	19
Link highlighting	5
Adjusting letter spacing	4
Reading mask	3
Reading guide	3
Easy-to-read text (font type change)	6
Sign language interpreter	3

Grouping elements is one of the factors that can influence the effectiveness of their utilization. Tools are stored in a separate panel, which can be accessed via a button, thus increasing the free space on the page. This approach requires the creation of additional symbols – icons that give access to the menu with accessibility tools. The above solution has been implemented in 9 university websites, such as the University of Life Sciences in Poznan and Opole University of Technology. Such an approach was necessary due to the fact that the websites of these universities offer a wide range of accessibility tools – expanded panels have up to 10 different types of tools. However, some universities still limit themselves to only two accessibility elements, namely changing the size of the text and changing the contrast. An example of this could be the website of the University of Life Sciences in Wroclaw.

Intermediary buttons/icons that launch a drop-down menu or open a panel with accessibility tools usually include a graphic symbol representing a person in a wheelchair. Icons with the symbol of an eye or eyelid are also used, which are sometimes combined with the word Dostępność. The A+ symbol has been used on the websites of the University of Bialystok and the University of Life Sciences.

After analyzing the university's pages, it can be concluded that there is a wide variety of markings among the accessibility tools used. The font resizing element is dominated by the use of three, increasing letters A with optional additional characters, such as A+, A++, A-, A-- and A+, A-. Less commonly encountered designations include Ab- and Ab+. Similar in appearance markings are used for contrast change tools. The letter A symbol placed in squares with a background in one of three colors (black, yellow, white) and a circle divided into two

parts – dark and light – are the two most popular graphic forms used for this element.

Another considered issue regarding accessibility tools on websites was their placement. The problematic location of these tools can contribute to the difficulties that arise in finding them. Conducting a review of selected sites led to several findings (table 2). The typical location of these types of tools is the top part which is usually a section of the header. On the other hand, considering the horizontal location, it is usually the right side or, less frequently, the middle. The exception is the website of the University of Gdansk, where the accessibility panel is located on the left, just below the header, and the Gdynia Maritime University, where the tools are located in the footer.

Table 2. The arrangement of elements on university websites

Criterion	The number of implementations
Top right corner of page	14
Top left corner of page	2
Center of right edge of page	3
Bottom right corner of page	3
Bottom left corner of page	1

2.2. Research objects

The websites of four public universities were selected as the objects of the study: University of Life Sciences in Lublin, Poznan University of Technology, Wroclaw University and Karkonosze Academy of Applied Sciences. Their choice was dictated by the fact that they had a similar interface structure and that their information layer did not contain too much content. Their homepages, or rather their graphics, were used to develop prototypes of pages displayed using a specially developed application. The purpose of this application was to display the prototypes and measure task completion times. Research participants worked with one of two versions of the scenario, during which 4 prototypes of the university's pages were randomly displayed. Before presenting each page, a board with a short instruction containing the content of the braiding to be executed appeared. The two prepared versions (A and B) differed in the way of presenting the tools: the symbols used, the layout, the indirect or direct way of access. The application displayed the pages in a dynamic way, and participants were expected not only to find the required tool, but also, in some cases, to use the tool. The application responded to the user's actions – changing the font size or contrast.

The first prototype contained the same elements, but they were placed in different locations. Version A of the site provided users with tools in the upper right corner, while version B provided them in the lower central corner. The prototype of Site 2 in both versions (A and B) had the same location of the tools – the lower right corner, but different symbols were used in them. In the case of the next website, the tools were grouped together in both versions. n the A version of the site, access to them is direct, while in the B version it is indirect - when you press on a button/icon, a menu with accessibility items was expanded. A similar situation occurred in the last prototype: indirect (A) and direct (B) access. The pages differed in the location of the elements: top right for version B and top left for version A.

2.3. Research group

The research experiment involved 35 people who were Computer Science students of Lublin University of Technology, both at the undergraduate and graduate level, as well as students of a technical high school with a technical profile. The average age of the respondents was 22. There were 32 men among the participants. Due to the specifics of the experiment (A/B tests), the study group was divided into 2 parts: the first included 18 people, and the second 17.

2.4. Experiment

The task implementation involved performing by participants one of two scenarios, which consisted of 4 tasks. Prior to the experiment, each person was briefed on the purpose and conduct of the study and agreed to participate in it. After the study, the participants filled out a questionnaire in which they evaluated and expressed opinions in relation to the prototype sites and the ways in which accessibility tools were presented on these sites. Table 3 presents the instructions aimed at finding and using the appropriate tools. For both Scenarios A and B, the content of the tasks was identical. During the execution of each task, the time of its implementation was measured.

Table 3. Tasks to be performed by the respondents

Task number	The content of the task
1	Use the tool that sets the page background color to black and the font color to yellow
2	Increase the font size on the page
3	Adjust the page contrast to make the background dark
4	Increase the font size on the page

The experiment was carried out with a Gazepoint GP3 HD eye tracker (figure 1) recording the position of the research participants' point of view while interacting with the prototype sites. The iMotions 9.0 software was used to prepare and implement the experiment. It allowed designing the experiment, performing calibration, recording the course of each participant's research session - recording actions in the interface, and later analyzing the recordings.

Fig. 1. Experiment

After the eye tracking part of the experiment, the respondents completed a questionnaire consisting of eight questions/tasks (table 4). In the user opinion questions, the survey results were expressed on a five-point Likert scale. The gathered information from the participants allowed for the determination of preferences for using accessibility tools.

Table 4. Instructions and questions of the author's survey

No.	The content of the question
1	Determine the best placement of accessibility tools to change contrast and change font size on web pages.
2	Indicate the best font enlargement system on websites.
3	Indicate the best way to present accessibility tools on websites.
4	Indicate the graphic symbol that best represents the accessibility tools used on websites.
5	Will the use of a background containing the university's view make it more difficult to use accessibility tools?
6	Indicate the best way to present the search engine on the website.
7	Indicate the best location of the search engine on the web pages.
8	Indicate the best contrast change system on the website.

3. Results

After conducting the experiment, eye tracker data and survey results were obtained. The eye tracking data was processed in iMotions software. It allowed to generate heat maps and scanning paths and to obtain the values of eye tracking indicators determined in defined areas of interest.

3.1. Measurement of scenario execution times

One of the most commonly used metrics in interface usability studies is task completion time (figure 2). This indicator can be interpreted in such a way that the faster the commands were executed by the participants of the experiment, the higher the efficiency was, and thus the better the quality of the interface.

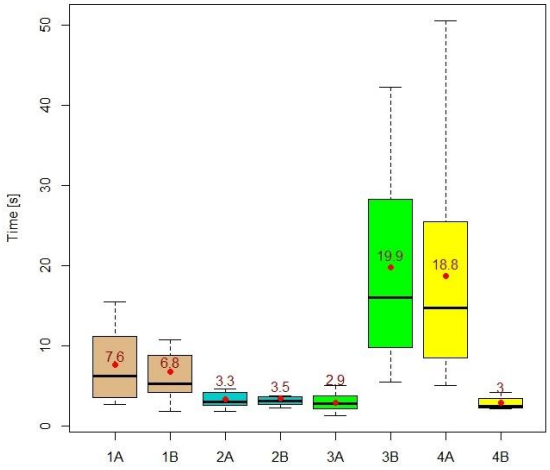


Fig. 2. Task completion times

For page 1, the average task completion times for the A and B versions of the page are comparable - the difference between them was small and amounted to 0.8 seconds. Reaching accessibility tools, normally located in the upper right corner, turned out to be slower than when they were located closer to the center of the page, below the large font text occupying a large area of the page.

For both versions of the second page, the times were almost the same. This was due to the fact that the elements searched for were located in the same place, i.e. in the lower right corner, and the tools differed only in the symbology used.

The execution times for the tasks in the two versions of the third page differed significantly. The difference was 17 seconds. Despite the same location (top right corner of the page), hiding the tool behind the icon and indirect access to the menu with accessibility items proved to be a major challenge for users.
A similar situation occurred in the case of the fourth page – here, too, a significant difference was observed in the execution times for versions A and B. Despite the same location of the tools (top left corner), the use of access via an icon, which, when clicked, expands a menu with tools resulted in a significant increase in the completion of this task. The reason for this can be attributed to the fact that there are additional actions that users need to perform during the execution of this task. These include moving the mouse cursor over the icon and then clicking on it with the mouse cursor.

3.2. Qualitative analysis

Qualitative analysis of the eye tracker results was conducted based on heat maps and scanning paths generated with the iMotions platform. Their preparation required time-consuming analysis of the recordings captured during the tests – marking in each recording the beginnings and ends of the display of successive stimuli. Heatmaps are graphical representations of fixation distribution superimposed on a static stimulus (page prototype). Scan paths, on the other hand, depict the individual activity of people watching or performing an assigned task on the displayed stimulus over time. Scanning paths take the form of different-sized circles representing fixations connected by lines, or saccades, and are superimposed on the displayed static image.

Figure 3 shows two heat maps – the result of visual scanning by the participants. In the case of version A of the site prototype, you can see the maximum concentration oriented in the upper right corner - where the accessibility tools are located. In version B, on the other hand, the hot spot is related to the location of accessibility icons positioned slightly lower from the center of the screen. In addition, the upper right corner of the page is also visibly greened, which is the place where respondents intuitively directed their attention first. This situation is illustrated by the example in figure 4, showing the scanning path for one research participant.

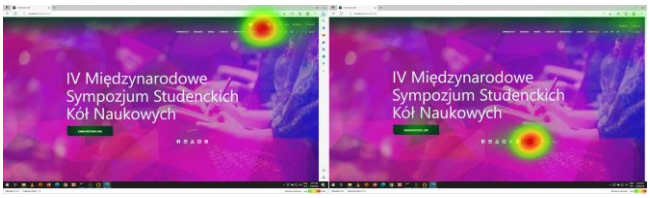


Fig. 3. Heat maps for the A and B versions of page 1

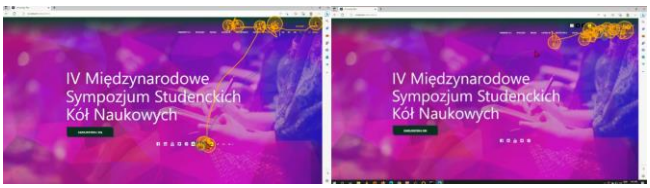


Fig. 4. Scan paths for the A and B versions of page 1

The second example (figure 5) relates to the fourth stimulus, in which the A version of the site includes an icon that allows access to accessibility tools by expanding the list. While the hot area clearly indicates the correct location, there are other places besides it that participants explored before making the final decision that behind the icon with the person in a wheelchair was an accessibility tool to enlarge the font. In the case of Scenario B, also visible is the hot area where respondents identified the target. However, in addition to this red-yellow-green circular area, only one warm green spot is visible - the upper left corner, to which the participants most likely intuitively and first directed their gaze. This conjecture is confirmed by the scanning paths presented in figure 6.

Fig. 5. Heat maps for the A and B versions of page 4

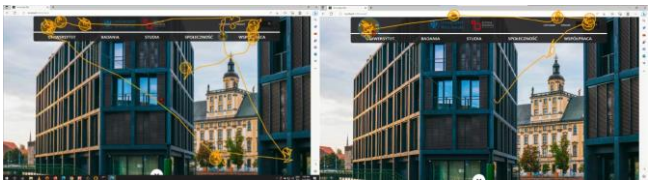


Fig. 6. Scan paths for the A and B versions of page 4

3.3. Quantitative analysis based on eye tracking indicators

The following eye tracking metrics were used for quantitative analysis: average time to first fixation, average number of fixations and average fixation duration. Each of the above metrics relates to a specially designated area of interest (AOI) covering the surroundings of a specific accessibility tool. Figure 7 summarizes the values of times to first fixation in the defined AOI.

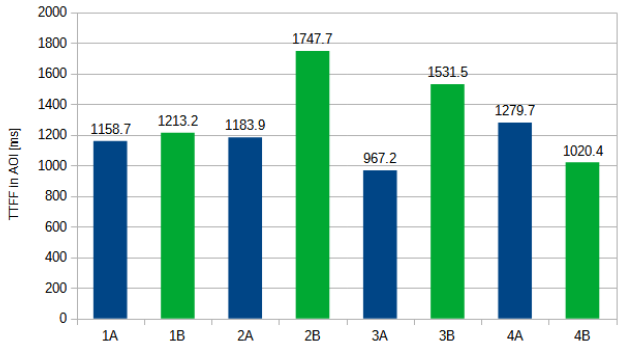


Fig. 7. Average time to first fixation in individual AOIs

Times to first fixation in areas of interest ranged from 967.2 ms to 1747.7 ms. For page 1 in both versions, the times were very similar because the icons searched for had the same distinctive appearance, but a different location. In version A, the accessibility icon was located in the upper right corner – the place where accessibility tools are most often located. The other icon was located in close proximity to the center of the page. In the case of page 2, the task completion times were clearly different, although the location of the tools was almost the same – the lower right corner. Additionally, the icons differed slightly in symbolism. The heat maps for the B version of the page show that the participants also observed other locations and elements of the page on their way to the target, such as the upper right corner and the icons for changing the language. This relatively large difference in average times to first fixation is puzzling in that the task completion times are similar. On page 3, the average time to first fixation was longer for version B, where the contrast change icon was accessed indirectly by pressing the icon with the wheelchair graphic, expanding the menu, and after selecting and clicking on the appropriate option. A similar situation occurred with prototype number 4. Here, too, the average time to first fixation was longer in the situation of indirect access to the target icon.

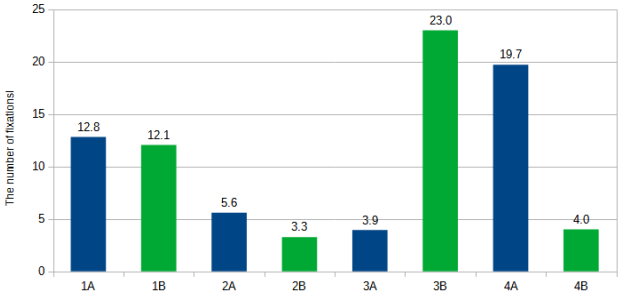


Fig. 8. Average number of fixations in individual AOIs

The next eye tracking indicator, the results of which are shown in figure 8, is the average number of fixations inside the areas of interest defined for each task and each version of the site prototype. The interpretation of this measure is that a higher average number of fixations takes place in those AOIs that include non-informative visual objects. For the prototypes of pages 1 and 2, the levels of this indicator for the A and B versions of the pages did not differ much. In the case of prototype 3, the high average number of fixations on the B version of the page was due to the fact that, on one side, the icon with the person in the wheelchair did not represent the contrast resizing tool but was used to expand the list with options, on which the contrast resizing tool was located. A similar situation occurred on page 4. A higher average number of fixations was encountered when the text resizing tool could be accessed by expanding the list. The higher value of this indicator may also have been influenced by the use of unusual symbolism in the form of tT letters.

The average fixation duration in AOI refers to one fixation (figure 9). This indicator may mean that the information in this area is particularly easy to read and understand, so participants do not need much time to process it. A long fixation time may indicate a high level of interest or task relevance in the context of the AOI.

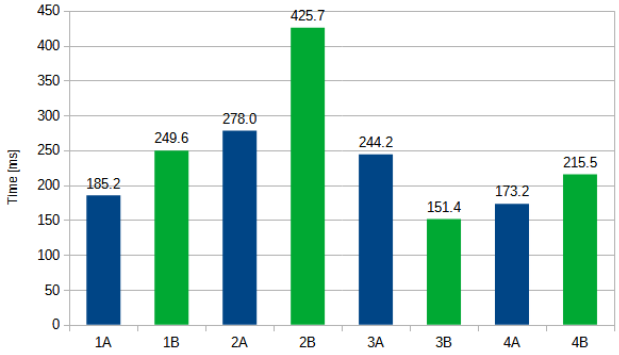


Fig. 9. Average fixation duration in AOI

For page 1, the B version of the prototype contained more icons per line, so participants needed more time to process the complex visual information. The A and B versions of the page 2 prototypes differed in their symbols. In the A version, the average fixation time was shorter because users intuitively chose the largest of the three A letters to reflect the enlarged text than it was in the B version, which used A-|A+ symbols. The longer average fixation duration for the A version of page 3 may indicate the greater relevance of this AOI. The white and black circle is a common symbol used to change contrast, while the wheelchair person icon is not used to change contrast. The results of this measure for the prototype of page 4 are analogous to those for page 3. Here, the same symbolism of a circle with a black and white background was also used, and therefore the average fixation time in this area was longer (version B) than on the tT icon in version A of the page.

3.4. Survey results

In the survey, respondents almost unanimously stated that the standard location for accessibility tools is the top-right corner of the page (figure 10).

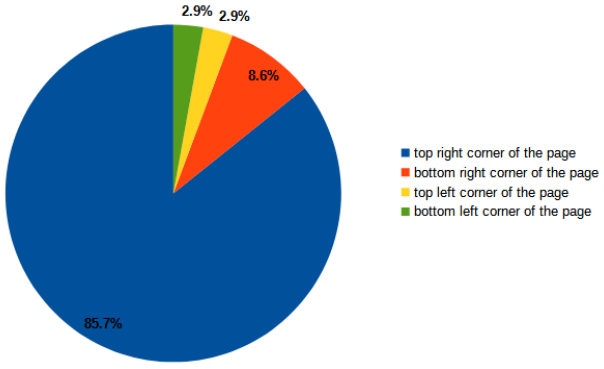


Fig. 10. Results regarding the location of accessibility tools

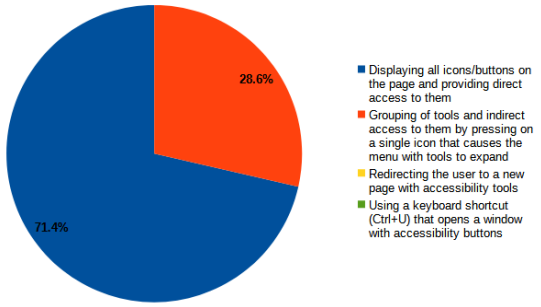


Fig. 11. Results regarding the method of accessing accessibility tools

According to the majority of respondents, when it comes to the method of accessing the elements, it should be direct – by displaying the appropriate graphic symbols in the page area (figure 11). In addition, 28.6% of respondents favored indirect access via a button/icon.

Figures 12 and 13 show the distribution of responses regarding the graphical presentation of accessibility tools. For contrast change items (figure 12), the most intuitive designation according to respondents is the letter A located in a yellow, black or white square. For changing the font size (figure 13), the options associated with combinations of the letter A combined with plus and minus signs were mostly selected. The button representing indirect access to accessibility elements, according to most respondents, should be combined with the word Dostępność in addition to the A+ or eye symbol.

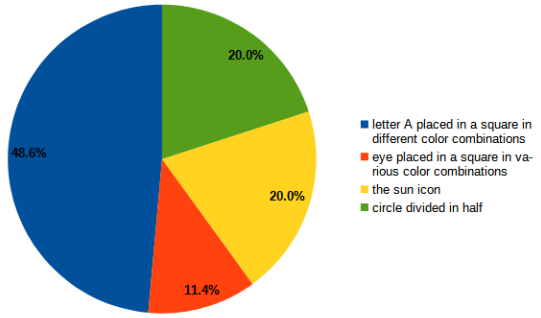


Fig. 12. Results of respondents' opinions regarding the choice of symbol for contrast change

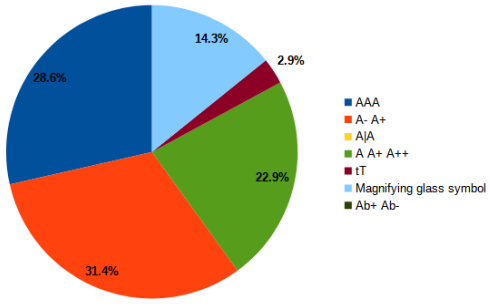


Fig. 13. Results of respondents' opinions regarding the choice of icon symbol for font size change

4. Conclusions

The purpose of the study was to verify the presentation of accessibility tools and its impact on the effectiveness of users working with these tools. On the basis of the collected data, from the experiment conducted using the eye tracking technique and the author's questionnaire, an analysis was carried out, which made it possible to formulate some conclusions and verify the research hypotheses. It turned out that various aspects of the graphical interface of websites, such a the placement of tools, the icon symbolism used and the way of accessing accessibility elements have an impact on the efficiency of people using them. Placing tools in an area of the site other than the top right corner increases the time it takes to locate them. When users enter a website, they first direct their gaze to this place, and only later search the other areas of the site. Another factor that increases the interaction time is the use of indirect access to tools. A single icon representing a button to a hidden panel or a drop-down list containing accessibility tools is more difficult to identify and find than when these tools are clearly located in one place close to each other. On the other hand, too many elements make the user have to look longer to reach a specific tool to find first and then recognize the symbol representing that tool. According to users, for changing the font size, the most intuitive solution is to use the letter A in combination with a plus or minus sign. However, the best way to reach the contrast change tool is to use an icon symbolizing a square with different background colors and the letter A placed inside it. On the other hand, the most effective form of search engine presentation according to respondents is a text field with the word Szukaj and a magnifying glass icon.

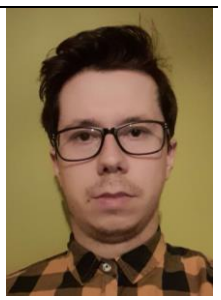
References

[1] Akram M., Bt Sulaiman R.: An Empirical Study to Evaluate the Accessibility of Arabic Websites by Low Vision Users. 8th International Conference on Information Technology and Multimedia (ICIMU), 2020, 206–211.
[2] Algül Y.: Web Accessibility of MOOCS for elderly students: the case of Turkey. Journal of life economics V(4), 2018.
[3] Ehmke C., Wilson S.: Identifying web usability problems from eye-tracking data. 21st British HCI Group Annual Conference on People and Computers: HCI...but not as we know it – Volume 1, 2007, 119–128.
[4] Esmeria G. J., Seva R. R.: Web Usability: A Literature Review. DLSU Research Congress, 2017.
[5] Insfran E., Fernandez A.: A Systematic Review of Usability Evaluation in Web Development. Web Information Systems Engineering – WISE 2008. Springer Berlin Heidelberg, 2008.
[6] Ma H., Zhao H.: Construction of High-Availability Teaching Website. International Conference on Management and Service Science, 2010, 1–4.
[7] Martin-Hammond A., Patil U., Tandukar B.: A Case for Making Web Accessibility Guidelines Accessible: Older Adult Content Creators and Web Accessibility Planning. 23rd International ACM SIGACCESS Conference on Computers and Accessibility, 2021, 1–6.
[8] Matera M., Rizzo F., Carughi G. T.: Web Usability: Principles and Evaluation Methods. Mendes E., Mosley N. (eds): Web Engineering. Springer, Berlin, Heidelberg, 2006, 143–180.
[9] Rodrigues S. S., Scuracchio P. E., de Mattos Fortes R. P.: A support to evaluate web accessibility and usability issues for older adults. 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, 2018, 97–103.
[10] Thoma V., Dodd J.: Web Usability and Eyetracking. Klein C., Ettinger U. (eds): Eye Movement Research. Studies in Neuroscience, Psychology and Behavioral Economics. Springer, Cham, 2019, 883–927.
[11] Țichindelean M. et al.: A Comparative Eye Tracking Study of Usability-Towards Sustainable Web Design. Sustainability 13, 2021, 10415 [http://doi.org/10.3390/su131810415].
[12] Vieritz H., Schilberg D., Jeschke S.: Merging Web Accessibility and Usability by Patterns. Computers Helping People with Special Needs. 12th International Conference ICCHP 2010. Procedings 12. Springer Berlin Heidelberg, 2010, 336–342.
[13] Web Content Accessibility Guidelines (WCAG) 2.1. https://www.w3.org/Translations/WCAG21-pl/ (available: 29.11.2022).

M.Sc. Marcin Cieśla
e-mail: marcin.ciesla@pollub.edu.pl

Graduate of the Lublin University of Technology in the field of Computer Science. His research interests include web development and, in particular, JavaScript-based technologies.

http://orcid.org/0009-0004-5736-5948



Ph.D. Mariusz Dzieńkowski
e-mail: m.dzienkowski@pollub.pl

Assistant professor in the Computer Science Department at the Faculty of Electrical Engineering and Computer Science at Lublin University of Technology. His scientific interests include human-computer interaction, eye tracking applications and web application development.

http://orcid.org/0000-0002-1932-297X

http://doi.org/10.35784/iapgos.3734 *received: 12.06.2023 | revised: 19.09.2023 | accepted: 14.11.2023 | available online: 20.12.2023*

INTERACTION METHOD BETWEEN WEBVIEW OBJECTS IN HYBRID JAVA APPLICATIONS

Denys Ratov, Oleh Zakhozhai
Volodymyr Dahl East Ukrainian University, Faculty of Information Technology and Electronics, Department of Information Technology and Programming, Kyiv, Ukraine

Abstract. *The article deals with method of interaction between JavaScript objects of different web pages in hybrid java applications. To solve this problem, the WebView component is used, its functionality to provide an interface for JavaScript objects, as well as the SharedPreferences global settings class, with its ability to store data in local storage. The software implementation is considered and the results of the practical use of the developed method of interaction between JavaScript WebView objects in the form of a hybrid Java application embedded in the electronic university system – the Timetable SNU electronic timetable module are presented.*

Keywords: Java, hybrid application, JavaScript, interaction, WebView

METODA INTERAKCJI POMIĘDZY OBIEKTAMI WEBVIEW W HYBRYDOWYCH APLIKACJACH JAVA

Streszczenie. *Artykuł dotyczy sposobu interakcji między obiektami JavaScript różnych stron internetowych w hybrydowych aplikacjach Java. Do rozwiązania tego problemu wykorzystywany jest komponent WebView, którego funkcjonalność zapewnia interfejs dla obiektów JavaScript, a także klasa ustawień globalnych SharedPreferences z możliwością przechowywania danych w lokalnym magazynie. Rozważono implementację oprogramowania oraz przedstawiono wyniki praktycznego wykorzystania opracowanej metody interakcji między obiektami JavaScript WebView w postaci hybrydowej aplikacji Java osadzonej w elektronicznym systemie uczelni – module elektronicznego planu lekcji Timetable SNU.*

Słowa kluczowe: Java, aplikacja hybrydowa, JavaScript, interakcja, WebView

Introduction

In today's digital world, where smartphones have become an integral part of our daily lives, mobile applications are becoming increasingly popular. This is especially true in the university environment, where effective scheduling is critical for students and faculty. Mobile applications allow students and teachers to quickly and conveniently view the class schedule without the need to use a computer or view a paper schedule. This brings convenience and flexibility to users, allowing for more efficient time and resource management. Using the mobile application, the class schedule can be updated in real time. This is especially important in case of changes in schedule or audience information. Students and teachers can quickly get updated information without having to check email. In this regard, writing a mobile application for the schedule of classes at the university becomes an urgent task [23].

Writing cross-platform applications requires a serious approach to planning future functions and capabilities of the program, since each of the selected target platforms [20], both theoretically and in practice, can impose its own limitations due to both the features of the target platform architecture and additional restrictions put forward by the developer company or the company that owns this platform.

Creating apps for Android is a complex process that requires detailed planning and has its own characteristics [18, 19]. One of them is the variety of mobile devices on which the application will run. When developing, it is necessary to take into account that users have different technical characteristics of their devices, and also use different versions of the Android operating system.

Today, if not all, then most Android devices have access to the Internet. Therefore, a large number of mobile applications, one way or another, interact with the Internet environment: they download files, log in and receive information from external web services, etc.

Hybrid applications have gained great popularity in the development of software for devices with the Android operating system. They combine the properties of both native and web applications [15, 16]. As a native app, it can be distributed to users through the app store, and it can also take advantage of numerous mobile device features. As a web application, it consists of HTML, CSS and Javascript files [17].

The benefits of this type of application include:

- many applications and interactive components can be written in JavaScript [21] for all mobile platforms;
- applications can use mobile device features such as camera, accelerometer, and others;
- all HTML, CSS and JavaScript files can be updated without waiting for a new version of the application to be approved.

To work with the network in Android, it is possible to use several methods. To get data from a specific Internet resource, you can use classes HttpUrlConnection (for HTTP protocol) and HttpsUrlConnection (for HTTPS protocol) from the Java Standard Library.

When developing a hybrid application to support the concept of Model-View-Controller [11], it is necessary to respect the separation of application data and control logic into three separate components: model, view and controller. The data received during the application from the activity web page is stored in the content model (JavaScript objects) of this web page [4]. Therefore, when loading other web content into the same activity, data is lost from the previous content model. So, for example, having loaded the content of a web page with authorization into an activity, it becomes necessary to transfer the result of the obtained authorization parameters to an activity with new web content. Therefore, there is a need to create a method of interaction between JavaScript objects of different web pages on activity in hybrid Java applications.

1. Application architecture

In the process of creating a hybrid application, among the standard elements, an important role is played by the WebView component [3], which is a full-fledged browser implemented as a subclass of View and capable of loading content from a specific URL, and therefore is designed to render html code. The operation of the component is based on the free engine for displaying WebKit web pages, which was developed by Apple. Using the WebKit engine ensures that the content will be displayed in the same way as in other browsers built on this engine – Google Chrome and Safari. Thanks to this, WebView can be used as a custom web browser, viewing content from the Internet through it [10].

The WebView component has many properties and methods that allow you to create the full functionality of a regular browser, and makes it possible to bind JavaScript code to Java code using an interface class [20]. To do this, the WebView component program uses the addJavaScriptInterface method, which is passed a class that provides an interface for JavaScript, and a name that will be used to display the instance in JavaScript (for example,

"iAndroid"). In the MVC concept, this ensures the creation of a controller component.

At the beginning of the application, the WebView component is initialized in the MainActivity class, using the identifier previously defined in activity_main.xml (Fig. 1). Then, to enable the existing WebView element to be associated with JavaScript objects, we set the interface class to the WebView [14].

```
browser  = (WebView) findViewById(R.id.webBrowser);
browser.addJavascriptInterface(new WebAppInterface(this),
                               "iAndroid");
```

Fig. 1. Setting the interface class in WebView

This will create an interface named iAndroid that will be available to JavaScript objects running in the content of the page loaded in the WebView component. With the help of the added interface class, the following possibilities appear:

1) call from the JavaScript module to the method described in the Java code;
2) execute a method from Java code, which is described in the JavaScript module.

Let's describe the WebAppInterface interface class with methods, the functionality of which is necessary in the JavaScript module code (Fig. 2).

```
public class WebAppInterface {
    private Context mContext;
    WebAppInterface(Context context) {   mContext = context; }
    @JavascriptInterface
    public void setParam(String key, String val) {
        String lastVal = "";
        prefEditor = settings.edit();
        if ( val.length() > 0 && ( key.equals("TM_idGroup") ||
             key.equals("TM_group") ) )
        {
            lastVal = settings.getString(key, "");
            prefEditor.putString(key,
                (lastVal.length() > 0
                    ? lastVal + ", "
                    : "") + val);
        }
        else {
            prefEditor.putString(key, val);
        }
        prefEditor.apply();

        if (key.equals("TM_FioUser")){
            Toast.makeText(mContext, (val.length() > 0
                                ? " збережено "
                                : " видалено ") + val,
                           Toast.LENGTH_SHORT).show();
        }
    }
    @JavascriptInterface
    public String getParam(String key){
        return settings.getString(key, "");
    }
}
```

Fig. 2. Interface class WebAppInterface

On the content side of the page loaded in the WebView, to interact with JavaScript objects with an instance of the Java class, we use the name of the created interface "iAndroid".

```
System.saveUser = function(){
let iStd    = $('#iStd'),    iTch    = $('#iTch'),
    inpFio  = $('#inpFio'), selFio  = $('#selFio');

    iAndroid.setParam('TM_sigUser', iStd.checked ? 's' : 't');
    if (iStd.checked){
        iAndroid.setParam('TM_FioUser', inpFio.value);
        iAndroid.setParam('TM_idGroup',
            selFio.options[selFio.selectedIndex].dataset.id_group);
        iAndroid.setParam('TM_group',
            selFio.options[selFio.selectedIndex].dataset.group);
    }
    else {
        iAndroid.setParam('TM_FioUser',
            selFio.options[selFio.selectedIndex].dataset.fio);
        iAndroid.setParam('TM_dolTeacher',
            selFio.options[selFio.selectedIndex].dataset.dol);
        iAndroid.setParam('TM_idUser',
            selFio.options[selFio.selectedIndex].dataset.id_fio);
    }
}
```

Fig. 3. An example of using the setParam method in the iAndroid interface

To pass parameters from web content objects to the iAndroid interface, we use its setParam method (Fig. 3). The name of the key and its value are passed as parameters. For reading, we use the getParam method (Fig. 4) with the key name parameter.

To maintain the separation of application data and control logic in the iAndroid interface, it becomes necessary to store data obtained from web content objects for further use. For example, user data, configuration settings, etc.

```
window.addEventListener( 'load', ()=>{
    System.status       = iAndroid.getParam('TM_sigUser');
    System.idfio        = iAndroid.getParam('TM_idUser');
    System.fio          = iAndroid.getParam('TM_FioUser');
    System.dol          = iAndroid.getParam('TM_dolTeacher');
    System.isOpenPanel  = iAndroid.getParam('TM_isOpenPanel');
})
```

Fig. 4. An example of using the getParam method in the iAndroid interface

To do this, Android has the concept of Preferences or settings. Settings are a group of key-value pairs that are used by an application. Settings are stored in xml files in unencrypted form in local storage. They are invisible, so they are not available to the average user.

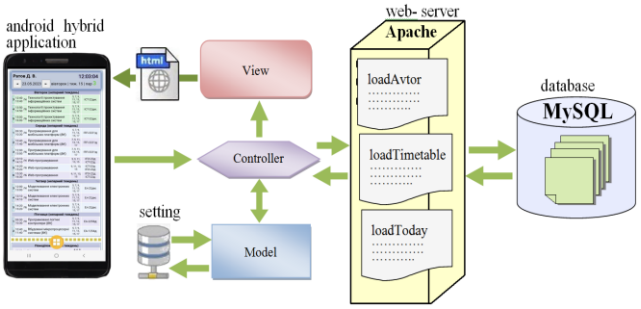


Fig. 5. Scheme of organizing support for the MVC concept in the context of a hybrid application

To store data from the content model, we use the android.content.SharedPreferences class. To manage the settings, an object of the SharedPreferences.Editor class is used. Asynchronous data storage allows key-value data to be stored on the mobile device, ensuring that the data is persistent even after the application is closed. The mechanism works asynchronously, which avoids blocking the user interface when saving or retrieving data. The use of asynchronous storage ensures the creation of model components in the MVC concept (Fig. 5).

A diagram of the process of internal interaction between JavaScript objects, an instance of the iAndroid interface class, and an instance of the SharedPreferences class is shown in Fig. 6. To make application-level requests to the server from JavaScript objects, the XMLHttpRequest API is used, which makes it possible to create asynchronous AJAX requests [2]. At the same time, the WebView component can reload its content using the loadUrl method.

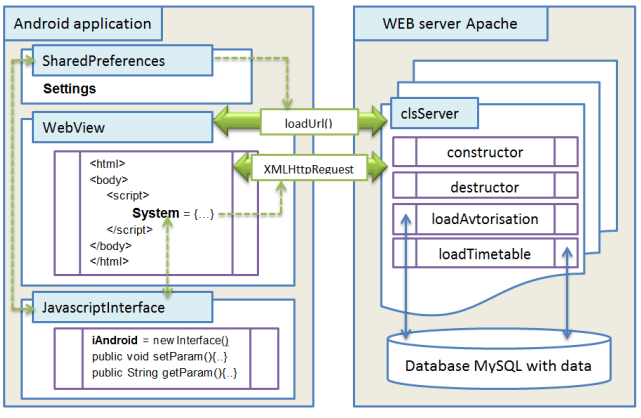


Fig. 6. Diagram of the interaction between JavaScript objects, an instance of the iAndroid interface class, and an instance of the SharedPreferences class

When forming a request to the server from the WebView component or JavaScript objects of any loaded content, the local storage (an instance of the SharedPreferences class) is first accessed. This makes it possible to get or save parameter values not only for cross-platform interaction between JavaScript objects and Java classes, but also between JavaScript objects of different content that is loaded into the same WebView component.

2. Results of the implementation of the method

The considered method of data exchange between WebView objects in hybrid Java applications has been tested in practice. It is implemented in the system of the electronic university, namely the electronic timetable module Timetable SNU – Volodymyr Dahl East Ukrainian University (Kyiv).

The application is complex, that is, it follows the client-server architecture pattern [1], which is one of the software architectural patterns and has become the dominant concept in the creation of distributed network applications and involves interaction and data exchange between them.

The server part of the schedule application of the busy university is responsible for the processes:

- Data management: saving and managing the schedule, information about classes, groups of students, teachers and other data.
- Data exchange via API: providing an API (application programming interface) that allows the client part of the application to receive and send data. This may include requests to receive a class schedule, update class data, or interact with other features of the application.
- Security and data protection: a centralized server part implies better data security control than a distributed system, where each element can have its own security flaws.
- Scalability and performance: the server simultaneously processes many requests and quickly responds to them, and also has mechanisms for caching and optimizing requests to the database. Scalability also includes the ability to easily expand the server infrastructure, for example, adding additional servers or using cloud services to ensure the reliability and availability of the application. This makes it possible to ensure the operation of the application when the load and the number of users increase.

The client part of the application implements software functionality:

- User interface: the application has a convenient interface that allows students and teachers to easily view the busy schedule by selecting a group or teacher through search, view busy details and changes. It is important that the interface is intuitive and provides convenient navigation through the application.
- Local storage of settings: the application can save the user's settings and personal parameters locally on the device. This allows you to store a user object, preferred settings and other personalized data without the need to constantly request them from the server.
- Interaction with the server: the application interacts with the server through the API, receiving data about the schedule, lists of teachers and groups, and other additional information.
- Mobile capabilities: the application can use the device's mobile capabilities, such as geolocation, camera, notifications, and others. For example, students can be given the opportunity to view the location of classrooms on a map, and teachers – the opportunity to download materials for each lesson.
  Local data processing: the application can perform some data processing locally on the device without the need for constant access to the server. For example, it can filter the schedule according to certain criteria.

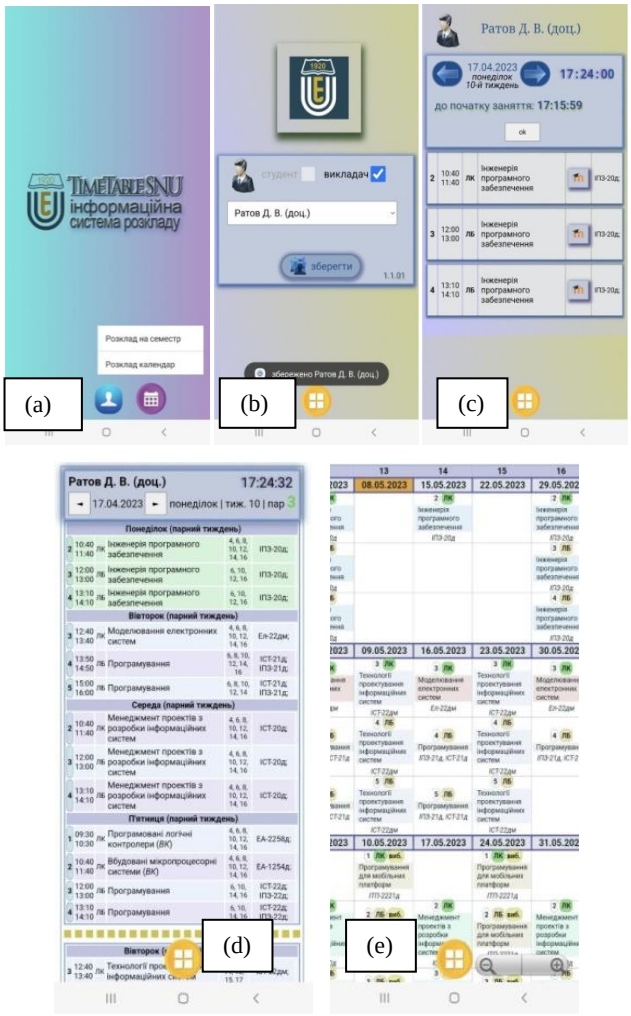


Fig. 7. Timetable SNU electronic timetable user interface

In terms of user interface (UI), the university class schedule application is divided into 2 Activities with data display that simplifies user interaction with the application (Fig. 7).

The main Activity contains application control components (Fig. 7(a)). The second Activity contains the WebView component, in which web pages of all functionality with different JavaScript objects are loaded from the server. These are the pages:

1) user authorization (Fig. 7(b));
2) a form with a schedule for the current or selected day (Fig. 7(c));
3) a form with a schedule for the entire semester (Fig. 7(d));
4) a form with a schedule in the form of a calendar (Fig. 7(e)).

The interaction between JavaScript objects of different pages loaded into the same WebView component is carried out through a chain of links: "iAndroid" – an instance of the interface class → "Settings" – an instance of the SharedPreferences class → "System" JavaScript object.

The program has created an access control system, which allows you to control access to resources and services, set access rights for different users or user groups. To do this, an object of the authorization and authentication system was created on the web page, which provides access only to authorized users. Having loaded the content of the web page with authorization into the activity (Fig. 7(b)), the received authorization parameters become available through the "Settings" JavaScript object to objects of other web content – the schedule page of the current day (Fig. 7(c)) and the schedule page for the semester (Fig. 7(d)). On the web page with the schedule form for the current or selected

day (Fig. 7(c)), JavaScript objects are created that execute asynchronous Ajax requests of the application level to the server. At the same time, the “Settings” object, an instance of the SharedPreferences class, participates in the requests.

Asynchronous Ajax requests of the application layer to the server allow students and teachers to connect to video conferencing platforms for remote classes. Also, the teacher has the opportunity to monitor their classes: the web page contains JavaScript objects that allow you to get a list of students who have attended the selected classes.

3. Conclusions

The method of data exchange between JavaScript objects in the WebView of hybrid Java applications described in the article makes it possible to get or save parameter values not only for cross-platform interaction between JavaScript objects and Java classes, but also between JavaScript objects of different content that is loaded into the same WebView component . This approach uses a modern and proven technology stack, which allows you to quickly develop and easily maintain the interaction of hybrid applications.

The method considered in the work made it possible to organize the interaction of JavaScript and Java objects in the developed Timetable SNU electronic timetable application. The application contributes to more efficient management of the class schedule at the university, which is critical for students and teachers. With the app, users can quickly and conveniently view class schedules on their mobile devices, eliminating the need to use a computer or view a paper timetable. This gives flexibility and efficient time and resource management. In addition, the chosen architecture allows you to always get up-to-date information about the schedule, which is especially important when the schedule changes. The mobile application allows you to get updated information quickly and conveniently without checking email or following ads. As a result of the introduction of a mobile timetable application at the university, students and teachers will be able to plan their time more efficiently, and the university will receive a convenient and modern tool for information support of the educational process.

The method of interaction between JavaScript objects and Java classes proposed in the work had a positive impact on the performance of the application's user interface and the possibility of scaling the functionality of the information system itself when implementing hybrid technologies in Java applications.

References

[1] Architecture Client-Server [https://ru.frwiki.wiki/wiki/Client-serveur] (available: 28.05.2020).
[2] Crane D., Pascarello E.: Ajax in action. Williams, Moscow 2006.
[3] Deitel P., Deitel H., Deitel E.: Android for Developers. Peter, St. Petersburg 2015.
[4] ECMAScript Language Specification – ECMA-262 Edition 5.1 [https://262.ecma-international.org/5.1/] (available: 28.05.2020).
[5] Expo – Create amazing apps that run everywhere [https://docs.expo.dev/get-started/expo-go/] (available: 28.05.2020).
[6] Expo Application Services (EAS) [https://expo.dev/eas] (available: 28.05.2020).
[7] Expo Go – Expo documentation [https://docs.expo.dev/get-started/expo-go/] (available: 28.05.2020).
[8] Griffiths D., Griffiths D.: Android Programming. O'Reilly Media, 2016.
[9] Kotlin Multiplatform – Kotlin Documentation [https://kotlinlang.org/docs/multiplatform.html] (available: 28.05.2020).
[10] Mednieks Z., Dornin L., Meike G. B.: Masumi Nakamura: Programming Android. O'Reilly Media, 2013.
[11] MVC architecture [http://www.gwtproject.org/articles/mvp-architecture.html] (available: 28. 05.2020).
[12] Niemeyer P.: Java Programming. Ekmo, Moscow 2014.
[13] Out-of-Tree Platforms – React Native [https://reactnative.dev/docs/out-of-tree-platforms] (available: 28.05.2020).
[14] Phillips B., Stuart K., Marsicano K.: Android. Programming for professionals. Peter, St. Petersburg 2017.
[15] Ratov D.: Architectural paradigm of the interactive interface module in the cloud technology model. Applied Computer Science 16(4), 2020, 48–55 [http://doi.org/10.23743/acs-2020-28].
[16] Ratov D.: Integration with the software interface of the com server for authorized user. Applied Computer Science 17(2), 2021, 5–13 [http://doi.org/10.23743/acs-2021-09].
[17] Ratov D.: Model of the user interface module of the information web system. Mathematical machines and systems 4, 2021, 74–81.
[18] React Native – Learn once, write anywhere [https://reactnative.dev/docs/getting-started] (available: 28.05.2020).
[19] React Navigation – Routing and navigation for Expo and React Native apps [reactnavigation.org] (available: 28.05.2020).
[20] Schildt H.: The Complete Guide. Williams, Moscow 2015.
[21] Stefanov S.:. JavaScript. Patterns. O'Reilly Media, 2010.
[22] vscode.dev – Visual Studio Code for the Web [https://code.visualstudio.com/blogs/2021/10/20/vscode-dev] (available: 28.05.2020).
[23] Zakhozhai O., Lyfar V., Ivanov V., Baturin O.: Uniform interaction model of educational process agents in unified management system of higher education institution. Information Technologies and Learning Tools 78(4), 2020, 266–277.

Ph.D. Denys Ratov
e-mail: denis831102@gmail.com

Assistant professor in Department of Information Technologies and Programming at the Faculty of Information Technologies and Electronics, Volodymyr Dahl East Ukrainian National University, Ukraine. The field of scientific interests is mathematical and computer modeling, decision support systems, software development using information technologies in applied fields and production.

http://orcid.org/0000-0003-4326-3030

Ph.D. Oleh Zakhozhai
e-mail: zakhozhay.oleg@gmail.com

Professor in Department of Information Technologies and Programming at the Faculty of Information Technologies and Electronics, Volodymyr Dahl East Ukrainian National University, Ukraine.
The author's area of research is focused on patterns recognition and data processing in complex systems, mathematical and computer modeling and visualization.

http://orcid.org/0000-0002-9078-3242

http://doi.org/10.35784/iapgos.5374 received: 24.07.2023 | revised: 18.08.2023 | accepted: 09.09.2023 | available online: 20.12.2023

BROWSERSPOT – A MULTIFUNCTIONAL TOOL FOR TESTING THE FRONT-END OF WEBSITES AND WEB APPLICATIONS

Szymon Binek[1,2], Jakub Góral[3]
[1]ClickRay Sp. z o.o., Cracow, Poland, [2]Kozminski University, Warsaw, Poland, [3]University of Economics, Cracow, Poland

Abstract. *The article presents the multifunctional BrowserSpot tool, which serves as an automated environment for testing websites and web applications for Android and iOS systems. It highlights and describes the individual stages of research and development work, the issues with solutions currently available on the market, as well as the project's results. The article also discusses the reasons for undertaking work on the tool, its functionalities, and the methods of its usage.*

Keywords: automation testing, bug tracking, smart test automation, responsive website

BROWSERSPOT – MULTIFUNKCJONALNE NARZĘDZIE DO TESTOWANIA FRONT-ENDU STRON INTERNETOWYCH ORAZ APLIKACJI SIECIOWYCH

Streszczenie. *W artykule zaprezentowano multifunkcyjne narzędzie BrowserSpot stanowiące zautomatyzowane środowisko do testowania stron internetowych oraz aplikacji webowych dla systemów Android i iOS. Wyróżnione i opisane zostały poszczególne etapy prac badawczo rozwojowych, problemy aktualnych rozwiązań dostępnych na rynku, a także rezultaty projektu. Przedstawiono również powody podjęcia się prac nad narzędziem, funkcjonalności narzędzia, oraz sposoby jego użytkowania.*

Słowa kluczowe: automatyzacja testowania, śledzenie błędów, inteligentna automatyzacja testów, responsywna strona internetowa

Introduction

This publication is the result of an R&D project carried out under a grant from the European Union between years 2017- 2019. The aim of our R&D project was to introduce a new "Software as a Service" (commonly known as SaaS) solution to the market. More specifically, it could be considered as a Testing as a Service platform [1]. It's main purpose is to automate the process of validation of website pages in the context of proper visibility on different devices and resolution types and generate automated reports regarding the results of the performed tests. We recognized the lack of modern automatic testing systems [3] that could work on testing different devices, process large amounts of data and also be a cloud based solution. The solution would also be much more user friendly thanks to the simple UX and UI design, codeless workflow and a drag&drop interface. The result of the R&D project was a fully working cloud based application, ready to be launched on the market for customers, as well as for corporate clients.

1. Literature and market review

During the start of our R&D project, there were already some solutions that offered automated tests of website pages but they either required technical knowledge to program and execute requests or they were not cloud-based. Since then, it has changed and we see more simulated solutions being present on the market. One of the industry's problems during the start of our R&D project was the lack of growth in noticeable increases in the efficiency and effectiveness of software testing. Most testing systems were designed on the basis of systems developed in the early 21st century. A big problem for the industry was the ever-increasing amount of data, making manual testing ineffective, and the increasing complexity of applications requiring more and more testing effort [4]. It was therefore necessary to redefine the work of testers, implementing new methods and practices.

The industry's problems were illustrated by the so-called Test Gap or the coverage testing gap (Fig. 1). It reveals areas that should have been tested, but could not be analysed due to time constraints, human resource limitations or high frequency of builds. An additional obstacle was that the testing gap is constantly widening, and the complexity of testing increases exponentially as new features are released. More code is produced than testers can test, leading to a gap in test coverage [2].

The main difference between the available solutions and our R&D project was that all of our tests would be performed on our cloud therefore we would not use the computing resources of our user's computers. In short, compared to the available solutions on the market, the user had to connect to a virtual machine and stay with the internet connection until the test finished. If the test failed for some reason, the automated test was stopped. A broken internet connection also caused the test to stop. Our solution, on the other hand, works based fully in the cloud, where the test scenario is created by the user using a drag&drop interface, and once finished, the test is sent to the cloud, where the algorithm distributes the test on the available devices. This also produces much better efficiency as compared to the available solution, which requires a constant connection to the internet and to the virtual machine according to the scheme one user-one virtual machine. The aim of the R&D project was to develop and prepare for market deployment a solution to address the challenges in test automation identified above, for which the name BrowserSpot was adopted.

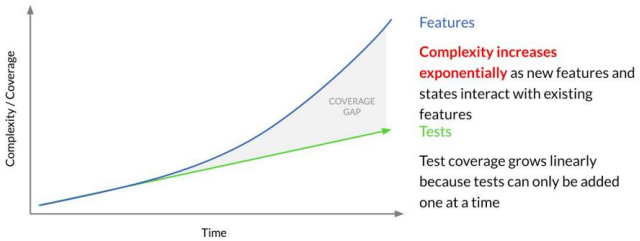


Fig. 1. The coverage testing gap theory [2]

2. Research and development

This chapter presents conducted research and development during the project. It was divided into eight stages that represent the goals and aims of individual stages of the project. Each stage was a combination of tasks related to both the research and development stages of this project.

First stage of the project focused on creating a code structure for the Selenium server, connecting it to browsers and devices. Key developments included algorithms for desktop and mobile browser operations, along with emulators for various platforms. The stage also aimed to verify technology functionality and browser compatibility with the Selenium server.

The aim of second stage was to research the possibility of effective separation of virtual environment technologies (created in stage 1) into physical machines and whether it will be possible to seamlessly communicate and exchange data

between the server and these machines in ranges of the hypotheses described in stage 1. The research will lead to knowledge of the technical possibilities of the server in conjunction with physical devices. A local environment will be created in which the technical capabilities of the server will be tested.

During the third stage of the project, research was carried out in terms of the introduction of cloud computing to the application. We also developed the user interface needed to run a SaaS solution at this stage. The conducted research gave an unequivocal answer to whether SaaS's cloud computing is a more optimal arrangement for the BrowserSpot solution. A report was produced to showcase the performance of the cloud base solution.

The fourth stage aimed to enhance the module's technical capabilities for improved site performance. This involved creating a component for page display performance, enabling the verification of page links, assessing rendering quality, error detection, and generating a detailed list of page element loading information. Key tasks included recording data to a database, enabling data export to the frontend, developing test generation capabilities, designing frontend components, implementing the project via API, testing module functionality, and creating and verifying necessary databases.

The fifth stage focused on expanding the capabilities of the language correctness testing module within BrowserSpot/component. This involved connecting it to external databases for research on grammatical correctness, stylistic accuracy, and content adaptation on websites. Activities included data handling, frontend integration, algorithm development, design, API implementation, testing, and database development.

The sixth stage aimed to enhance the module's capabilities for comparing elements on rendered web pages across different browsers. It involved developing a component to identify errors in individual page elements and generate descriptive reports about differences in their display. The module extracted data from the server and processed it to detect errors related to specific elements on web pages. Key tasks included algorithm development, test generation for error detection, element comparison, error presentation, integration with the frontend, and technology verification.

The seventh stage involved conducting industrial research to acquire knowledge and skills in data conversion, communication, and presentation. A component was developed to retrieve errors in various website elements and generate descriptive reports on differences in element display across different browsers. Research included developing a communication algorithm and protocol for handling website rendering data and presenting it in the user interface.

The technical solutions obtained would be part of a prototype for the BrowserSpot service. Goals for this stage included generating comprehensive reports combining correctness checks for page display, linguistic correctness, and individual element examination, as well as creating partial reports as notifications for user activities related to verification.

Activities included developing algorithms for data retrieval and export to the frontend, conducting tests on data processing and report generation, PDF report generation, frontend component design, API implementation, integration, and various tests to ensure module and tool correctness. A database algorithm was also developed for module operation.

The eighth stage focused on integrating all technology elements and conducting real-world testing for the BrowserSpot service pilot. The goal was to complete a prototype of the BrowserSpot tool and test it in real conditions, including demonstrations among a selected group of users.

Prior to the demonstrations, activities included integrating technological components, addressing errors in information generation and browser operations, improving the API, finalizing the design of components and the prototype, making frontend adjustments, and verifying the tool's operation in terms of error generation, API functionality, frontend performance, and overall workload.

The prototype demonstrations aimed to verify assumptions regarding performance, user support, and concurrent testing, the correctness of error generation, report generation in various software components, interface functionality, and identifying and resolving any performance bottlenecks in the tool.

3. Results

The aim of this chapter is to present results of each individual stage of the project in order to give an insight on which aims and goals have been properly achieved, and which were more problematic.

During the first stage of the project, a virtual machine was created as the runtime environment and the core of the entire project. A virtual machine is an isolated environment running on a host computer, and simulating the operation of a physical device with a separate operating system, freely chosen independently of the host computer system. The main component running in such a prepared environment was an engine based on Selenium technology. Selenium is a framework that automates tasks related to functional testing of web applications. It allows you to control web browsers from the code level and define sets of actions to execute within them. Selenium itself includes a set of tools - among them, WebDriver was used in the project. Its integration with Java was used. After the creation of a working prototype of the BrowserSpot tool engine with such features as: configuration and readiness of the virtual environment for further development, server startup, creation of an algorithm to unify support for different browsers, the ability to remotely operate browsers (Google Chrome, Mozilla Firefox, Opera, Safari, Internet Explorer) (Fig. 2) on the virtual machine, the ability to remotely operate browsers on mobile devices emulated within the virtual machine, logging of the course of action of browsers, the end of this stage have been reached.

After the first stage of work, the system used virtual machines supported by standard desktop operating systems and a set of mobile device emulators. The operation of the latter offered an approximation of the behavior of the devices, and thus the ability to test the operation of sites in their environments, but did not correspond 1:1 to real use cases due to a number of drawbacks that the use of emulators entails. So at this stage, as much of the emulators as possible were replaced with physical devices. This required designing proprietary mechanisms for integrating physical devices into the tool engine. The algorithm had to provide reliable two-way communication with the device. A solution was implemented to support physical devices from the desktop computer and display the screen content of the mobile device on the computer screen - the ability to perform real-time operations - latency was reduced to a minimum. A similar solution was applied to desktop systems. New functions were designed and implemented, such as downloading screenshots containing web pages rendered by different browsers, downloading codes of tested sites, comparing how sites are displayed by different environments. Implementation of new functionalities and development of existing ones allowed the creation of a database. The database was to be used to store all data necessary for the operation of the site and arising during its use. At this stage of the project's advancement, these included user data, logs, screenshot data, and test history. Due to the proliferation of system functions and the connection of new devices to the system, the need to implement load balancing methods was recognized. To this end, a load balancer was used, which is a solution for distributing tasks among devices in the infrastructure. To conclude the second stage of the project, a working prototype was developed with the following features: use of physical devices alongside emulators, ability to remotely operate physical devices in real time, downloading screenshots, comparing page renders, implemented and secured database, use of load balancer.

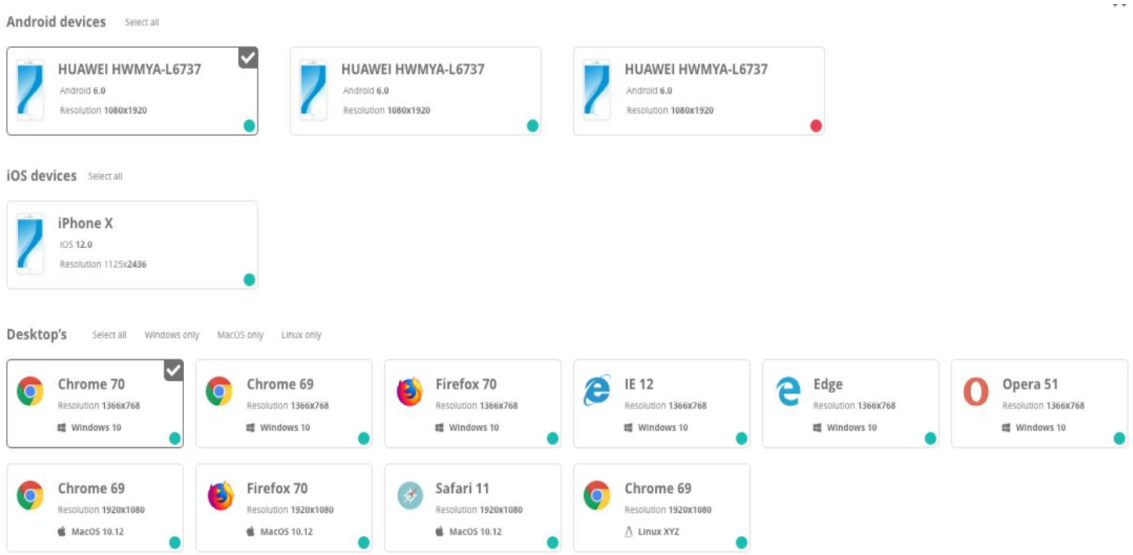


Fig. 2. Ability to run tests on multiple devices, browsers and systems using

Table 1. Traditional and SaaS model comparison

Traditional model (on premise)	SaaS model
• The user has to maintain the entire system manually in order to conduct tests (different browser versions, etc.). • The performance of the software (speed of conducting tests and task execution) is limited by the user's computer computational power. • Blocking other user activities while conducting tests (tests do not work in the background).	• The user receives the whole package, including the software and infrastructure (devices in the cloud with various browser versions, different resolutions, without incurring device maintenance costs). • The model can be scaled by enriching the cloud with new devices; a single task is distributed and executed simultaneously on multiple computers, and traffic can be balanced using a load balancer. • The user's computer is not burdened as the application natively launches browsers, and tests are executed in the background.

During the third stage, a performance comparison module has been developed between an application running in a traditional model (when the user installs the software on his computer) and software delivered as a service (SaaS model). The analysis was carried out from the perspective of user usability.

Performance calculations were carried out for a number of test cases. Estimated software performance (speed of testing and task execution) for an example test case – to test a page of 57 subpages, on 3 browsers. Assumptions used for calculations: average time to test 1 sub-page using one computer – 35 s (averaged value, the exact time depends on the parameters of the device such as the amount of RAM, processor speed, etc.). The results of our SaaS model showed a drastic improvement in performance. Depending on the complexity of the webpage and the components of a compared traditional model computer, we have noticed a performance increase of up to 70% using our SaaS model solution.

Given these results, the focus was on moving the service to the cloud and making it available in a SaaS model – Software as a Service. This is a cloud computing model in which a service is made available to users remotely over the Internet, and all computing is done on the server side. Until now the service operated locally, requiring installation and proper configuration by the user, or rather, the developer. The key advantages of moving the service to the SaaS model include no need for the user to install the software which results in a lower entry threshold for potential customers, no specific hardware requirements, easier distribution of the solution, easier maintenance of the service in the long term, increased scalability of the system. To enable the service to operate in this model, it was necessary to create a user interface. This part of the work included: establishing the image of the application, designing layouts, implementing layouts, designing and implementing mechanisms for controlling remote devices, integrating the front-end with the server's API, testing the interface, starting work on the ongoing maintenance of the service.

The completion of stages form 4 to 6 have been reached after the successful implementation and testing of: the performance testing module, the language testing module and the render comparison module. The automatic site quality (performance) testing module allows you to verify a website for performance and correct application of SEO practices, and suggests possible improvements. The test is carried out completely automatically – the user only enters the website address, and assigns a given test to: the client, the project and a milestone he/she has set. Following the performance test, a report is automatically generated that contains a range of data, including, in particular: information about the quantitative share of the server's individual http response codes (200, 301,404), information about the amount of data downloaded from the server along with detailed addresses from which these data were downloaded, information about the amount of data downloaded from the server taking into account the type of data (html, css, js, etc.), basic summary information about response and page load times, suggestions for site improvements (YSLOW) taking into account Javascript code optimization, file compression of the number of HTTP requests, etc.

The second functionality introduced during these stages was a module for testing the linguistic correctness of the site. Running on the basis of the LanguageTool tool, the algorithm checks all the text visible to the user and checks it for: grammar, spelling, phraseology, punctuation, syntax. After the test, a report is generated with suggestions for improvements. The third functionality launched at this stage of work was a module for comparing renders of different browsers. The tool is based on previously developed functionalities – downloading screenshots, html code and a tool looking for differences. When run, the module returns a result containing previews of the page rendered by each of the selected browsers. Each of them can be freely viewed and manually compared. The main element of the tool, however, is an automatic comparison of rendered pages. The tool allows you to: select two renders and juxtapose them against each other, synchronously scroll through the rendered pages for visual comparison, automatically find and list differences, set a tolerance for differences in element sizes, review and locate differences, search for elements that are missing, search for elements that appeared in the code, but you are not sure about their display on the user's side. In addition, we have developed component responsible for retrieving detected errors in the scope of individual elements of the studied website

and generating reports automatically presenting information in a descriptive manner on the differences in the display of individual elements of the rendered page on different browsers.

During the seventh stage of the project, a component responsible for generating reports for the end user was developed. Functionalities related to storing and sharing historical and current reports have been implemented. All reports are stored in a secured database. The repository contains reports from 3 types of tests: render tests, performance and SEO tests, and language tests. Each of the reports is available in the BrowserSpot tool itself under the "Reports" tab, as well as it is possible to download them in the form of a formatted PDF document for presentation outside the service or in paper form.

The eighth stage consisted of the merge of technology components developed in earlier stages, a series of tests, and a prototype of the BrowserSpot tool that was made available to a closed group of users for UX testing. As a result of the work carried out as part of the project, a complete IT product was created with the following features and functionalities: backend based on technologies: Java and Selenium, virtualization of devices in an isolated environment, encrypted database storing application and user data, support and functions for remote control of selected physical devices making up the device testing lab, control of network traffic with a load balancer, making the service available in the SaaS model, mechanisms for handling user accounts and workspaces, rendering web pages on emulated devices, virtual machines and physical devices, downloading rendered pages, downloading source code of pages, automatic comparison of renders between browsers, automatic analysis of text displayed by browsers for correctness and suggestions for improvements, automatic creation of reports on performance and meeting SEO standards organization of the above tests by client, ability to plan tests for the future and create a schedule, archiving of historical tests, creation of an API for the front-end interface, creation of a front-end interface, tutorials, access to documentation, modifiable interface of the management panel, support for automatic performance testing functions, SEO, language, renders, support for the report archive view, support for the schedule, insight into statistics. The described functionalities went through a cycle of tests and corrections to eliminate possible errors in both the application design and its implementation. Subsequently, UX testing was carried out in accordance with the following methodological assumptions. The results of the survey were collected using a dedicated questionnaire. The data source was a task test with a framework scenario - each participant was tasked with using BrowserSpot in a real - current or historical - case. The requirements for the task were formulated as follows: the task is related to the testing of any service, the object of the task is a web application with an average level of operational complexity, the task involves a single application or its component. The group of respondents was selected from among programmers, QA staff (software testers) and IT project managers. The surveys were conducted asynchronously, without a moderator. There were 3 areas of problem analysis: quality of solutions provided, redundancy or functional deficiencies, bugs and overall service performance. The study participants were provided with a manual and basic user materials, and no guidance or advice was given during the test. The study was anonymous, the only data of the participants provided during the course of the study was their position and seniority. The results obtained were used to improve the BrowserSpot service before its full rollout to the market.

4. Selected applications

4.1. Translation of the execution code

The code that is used to take a screenshot of the website has to be translated based on the device, software version and used application. The tool checks the selected parameters and does the translation in the background, without user input in order to ensure its ease of use. The translated code is then executed on the website and the results are compared and verified. Here are the examples of the translated code snippets for the use in the chrome browser.

The function to scroll to the top of the page and wait 2 seconds for the page to fully load:

```
    {
this.driver.executeScript("document.querySelector('body').scrollTop=" +
(0) + ";document.querySelector('html').scrollTop=" + (0) + "");
        try {
            Thread.sleep(2000);
    }
```

The function to take a screenshot of the page:

```
    {
TakesScreenshot takesScreenshot = (TakesScreenshot) new
Augmenter().augment(this.driver);
byte[] image = takesScreenshot.getScreenshotAs(OutputType.BYTES);
    }
```

4.2. Environment scalability

The tools environment has been developed in a way that allows an easy expansion if necessary. New devices, software, and versions can be added and quickly configured on the backend of the tool, but also the new systems and machines that handle the operation functions of the tool. That way the need for more computation power can be solved by simply adding a new server or computer and download the appropriate software that will automatically set itself up and keep updated with the rest of the environment.

4.3. Tasks load balancer

The tool has been designed with task load balancing in mind. The way this problem has been solved is by running a load check before sending a task request to the specific machine. If the check finds a machine that is currently under load it forwards the request to the next one and the process begins again. The simplified way is illustrated on the following chart (Fig 3).

4.4. Physical devices

The tests conducted using the tool are requested to be performed on a series of physical devices that are connected to the system. Usually these types of tests are performed on an emulator on virtual machine that simulates the selected device. The decision to use physical devices has been made mainly due to the numerous problems with emulators but also for the performance and accuracy of the tests and results.

5. Conclusions

As a result of the conducted R&D project, a comprehensive testing tool has been developed, allowing for significant automation and standardization of testing activities, particularly in small and medium projects. The results of individual project stages have verified the assumptions defined at the beginning of the R&D project. The results have been implemented in the market. The tool is offered under the commercial name BrowserSpot in a SaaS/TaaS model. The tool addresses key challenges faced by the automated testing industry, primarily related to technical infrastructure and high costs of maintaining device laboratories and human resources. The developed tool helps alleviate staffing issues and is accessible to individuals who do not possess extensive programming skills. After the completion of the R&D project, the tool was further developed to better align with user needs and expectations, such as enriching it with a function for creating codeless automated tests using a graphical user interface and drag-and-drop functionality and progress tracking dashboard (Fig. 4).

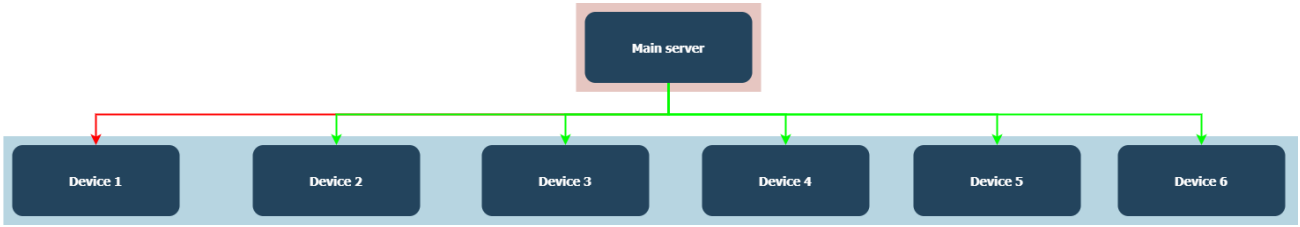


Fig. 3. Load balancing method flowchart

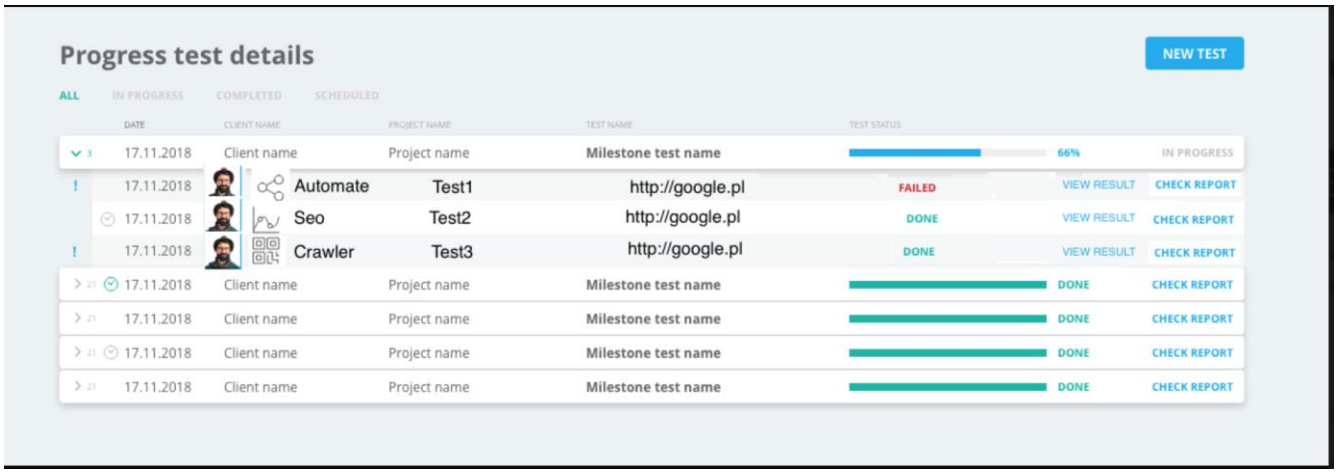


Fig. 4. The function of progress tracking dashboard

Acknowledgments

The R&D project, the results of which were presented in the article, was co-financed using a grant by the European Union. Project was titled „Developing a unique rendering mechanism to launch a proto-type BrowserSpot service". Research project no. RPMP.01.02.01-12-0487/16 co-financed from the funds of Regional Operational Programme for the Malopolska Region 2014-2020, Priority Axis Knowledge Economy, Measure 1.2 Research and innovation in Enterprises, Sub-measure 1.2.1 R&D projects of Enterprises co-financed by the European Regional Development Fund.

References

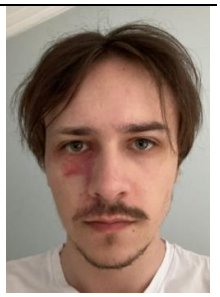
[1] Ali A. et al.: Automated Parallel GUI testing as a service for mobile applications. Journal of Software: Evolution and Process 30(10), 2018 [http://doi.org/10.1002/smr.1963].
[2] Arbon J.: AI for Software Testing. Pacific NW Software Quality Conference, Portland 2017.
[3] Moreira R. M. et al.: Pattern-based GUI testing: Bridging the gap between design and Quality Assurance. Software Testing, Verification and Reliability 27(3), 2017, e1629 [http://doi.org/10.1002/stvr.1629].
[4] World Quality Report. Capgemini, 2019 [http://www.capgemini.com/news/press-releases/world-quality-report/] (accessed 29 June 2023).

M.Sc. Szymon Binek
e-mail: s.binek@clickray.eu

He is the main originator and the co-founder of ClickRay. He is a specialist in online and HubSpot API development services and a truly wide range of digital ventures and experiments for clients. He supervised a research and development project co-financed by European funds, which resulted in the BrowserSpot platform.
Research interests: Artificial Intelligence and Machine Learning, Software Test Automation.

http://orcid.org/0009-0001-7936-8056

B.Sc. Jakub Góral
e-mail: goral.jakub99@gmail.com



In 2022 he received a Bachelor of Science degree (B.Sc.) at the University of Economics in Cracow in the field of Business Management with a specialization in Small Business Management. Later on he continued his studies and is now in the process of getting a Master's degree in International Business Management.
Research interests: technology, Industry 4.0, Artificial Intelligence, Data science.

http://orcid.org/0009-0003-9634-4915

http://doi.org/10.35784/iapgos.5376 *received: 26.07.2023 | revised: 19.10.2023 | accepted: 20.11.2023 | available online: 20.12.2023*

ADVERTISING BIDDING OPTIMIZATION BY TARGETING BASED ON SELF-LEARNING DATABASE

Roman Kvyetnyy[1], Yuriy Bunyak[2], Olga Sofina[1], Oleksandr Kaduk[1], Orken Mamyrbayev[3], Vladyslav Baklaiev[4], Bakhyt Yeraliyeva[5]

[1]Vinnytsia National Technical University, Vinnytsia, Ukraine, [2]Spilna Sprava Company, Vinnytsya, Ukraine, [3]Institute of Information and Computational Technologies of the Kazakh National Technical University named after K. I. Satbayev, Almaty, Kazakhstan, [4]Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, [5]M. Kh. Dulaty Taraz Regional University, Taraz, Kazakhstan

Abstract. *The method of targeting advertising on Internet sites based on a structured self-learning database is considered. The database accumulates data on previously accepted requests to display ads from a closed auction, data on participation in the auction and the results of displaying ads – the presence of a click and product installation. The base is structured by streams with features – site, place, price. Each such structural stream has statistical properties that are much simpler compared to the general ad impression stream, which makes it possible to predict the effectiveness of advertising. The selection of bidding requests only promising in terms of the result allows to reduce the cost of displaying advertising.*

Keywords: advertising bidding, targeting, targeted advertising, click prediction

OPTYMALIZACJA OFERT REKLAMOWYCH POPRZEZ UKIERUNKOWANIE W OPARCIU O SAMOUCZĄCĄ SIĘ BAZĘ DANYCH

Streszczenie. *Rozważono metodę ukierunkowywania reklam w serwisach internetowych w oparciu o ustrukturyzowaną samouczącą się bazę danych. W bazie gromadzone są dane o wcześniej zaakceptowanych żądaniach wyświetlenia reklam z zamkniętej aukcji, dane o udziale w aukcji oraz o wynikach wyświetlania reklam – zarejestrowanie kliknięcia i instalacji produktu. Bazę tworzą strumienie z cechami – strona, miejsce, cena. Każdy taki strumień strukturalny ma właściwości statystyczne, które są znacznie prostsze w porównaniu do ogólnego strumienia wyświetleń reklamy, co pozwala przewidywać skuteczność reklamy. Selekcja tylko obiecujących pod względem wyniku zapytań ofertowych pozwala na obniżenie kosztów wyświetlania reklam.*

Słowa kluczowe: licytowanie reklam, ukierunkowywanie, reklama ukierunkowana, przewidywanie kliknięć

Introduction

Internet bidding became very important factor of economics support on condition of the pandemic. The on-line bidding often starts from the moment when consumer saw an interesting product on a banner of some site. The consumer can make click on the banner and he will be redirected to the corresponding shop, where he may buy the product or not.

From technical side of view, this process is looking as bidding of banner places by internet providers for advertisers. The provider controls the moment of a consumer visit to site and sends bid requests (BR) message to advertisers to participate in a closed auction of advertising (AD) places on the site. Since the display of advertising is paid, the advertiser tries to participate in those auctions where his advertising campaign will be successful. Advertiser specifies by targeting methods the BR messages which most closely correspond to purposes of an advertising campaign and have highest probability to receive paid events – consumer's click event (CLICK) and event of buy by consumer of advertised product (INSTALL). In the Real Time Bidding (RTB) this is exchange between Demand Side Platform (DSP) and Supply Side Platform (SSP). The aim of the targeting is to reduce cost of the CLICK and INSTALL.

The first step of the targeting is known as Native Targeting (NT). It is accessible by the way of the definition native parameters, such as country, region, time intervals, AD categories, list of preferable sites, creatives size, language, user gender and year of birth, etc. All known applications for media buying provide the ability to set NT parameters. The aim of the NT is to specify optimal stream of bid request messages and to limit RTB server load. The NT is not responsible to made the targeting effective, because, less than a tenth part of the BR in the received optimized by NT stream are useful in terms of getting CLICK and less than thousandth part give INSTALL.

At a current time, there is only one approach to made AD trades effective, which is widely using by professional companies – this is the CLICK prediction. There are many methods of its implementation. The modern methods are basing on using of Data Management Platform (DMP) in the manner of real time database. The database stores a history of participation of the BR messages parameters values in previous trades and related with them results. When current BR was received the history of the previous results associated with parameters of the BR create its rating and help to make a decision about participation in the trade. The DMP also includes statistical models of events occurring against the background of the BR stream.

The AD auction creates a problem for the SSP associated with a choosing of a minimal price for the lot received in the BR message so to win the auction. The closed bidding is the feature of the auction. Therefore, only the SSP which won the auction lot knows the lot selling price, offered by nearest competitor. So, the RTB needs in price strategy for forecasting a minimal winning price by the help of a dynamic model which reflects price changes on current bidding conditions.

1. Materials and research methods

The mathematical background of the AD utility estimation and forecasting is wide – from Bayesian probability model to matrix and tensor models with implementation using convolutional and neural networks. The models are intended to evaluate and predict auction winning process, AD utility parameters such as Click to Bid Ratio or Click-through-Rate (CTR), Install-through-Rate (ITR) and other. The RTB process is described in detail in [1]. There are two specifications of messages used in bidding transactions [6, 16]. The BR messages defined by these specifications are similar and include some equivalent parameters and some specific parameters which sign user and device location, thematic profiles by differ ways. The Open RTB specification [4, 11] was used in [1, 3]. Authors made review of some approaches to forecasting of the RTB price and AD budget optimization. There are known probabilistic models which taking into account both the behavior of the users the advertisers, based on history of the impression, the time or date of the impression, the presence of social functionality. The aim of many authors efforts is to predict the bid value while acquiring an impression at a lowest cost. Their strategy is based on a win model which predicts the winning price based on a regression model. The authors of [1] have evaluated the feasibility of applying forecasting approach using autoregressive integrated moving average model to predict the bid prices. The accuracy of the predictions was very low due to dynamism of the RTB process. At the next step they developed

a dynamic programming algorithm to bid for the impressions that operates over a set of consecutive time periods. The algorithm follows a model which adjusts its properties for the next bid period based on the prior period behavior. The model adjusts bid price with account of budget strategy and reached wining results. Authors affirm that such approach can adapt the bidding process in the RTB successfully. The performance of the algorithm depends on duration of chosen bid period. The AD utility and its relation with bidding model was not considered.

The idea to take into account the value of the highest competing bid for prediction of the AD utility was considered in [4]. The utility was interpreted as a profit of the bidding. Expected utility is the integral of distribution of the highest competing bid on condition of bid sale price and corresponding them winning and losing events. The integral was evaluated by using a given set of historical events. So, the highest bid price is a factor of AD utility from the point of view of bidding profit.

Bayesian approach to the prediction problem yields the model of Logistic Regression (LR) [6,11] which is widely used for bidding utility prediction with account of data structures. The next value of the utility parameter, for example the CTR, can be evaluated by using the current and previous values vector and weighting vector which is estimating for each step on the condition that logarithm of click event and not click event probabilities relation (the loss function) is equal to scalar product of the vectors. The procedure of the weighting vector evaluation is a nonlinear optimization problem. Different ways of its iterative solution and implementation are discussed [2, 5, 9, 10]. But LR based methods cannot capture higher order interactions between features, which have proved to be important in the CTR prediction [5, 7]. Therefore, the LR is using in combination with data structures which reflect bidding process [14].

The factor model (FM) of the utility prediction was offered in [13]. The model is basing on pairwise interactions of ID sources. The ID are identifiers of advertisers. The pairwise interactions create a matrix. Every model matrix cell has its own set of features for the factorization. The final prediction is the sum of all pairwise feature dot products. The advantage of this approach is that the information from the test samples set is capable for predicting for new IDs.

The problem of CTR prediction by the DSP is to calculate the bid price according to the estimated CTR is presented in [14]. Tensor factorization model of bidding data and click events was offered. The model presents coupled interactions between user, publisher and advertiser as third order tensor. The tensor is sparse and therefore it can be effectively factorized using high order singular value decomposition. As the result, the CTR is estimated in a manner of a linear sum of some factors related with interaction model. The analogues sparse structures and their parameters optimization by clear and latent factors retrieving using the Method of Factorization Machines (MFM) are considered in [8, 12, 15]. The MFM is model class that combines the advantages of Support Vector Machines with factorization models. The model presents the click probability or the loss function as the second order nonlinear function of a feature vector of impressions. The field-aware version of the MFM [8, 17] is presented by released an open source software.

2. Model experiment

As it follows from above short overview, the RTB utility is an object of influence of many factors. The factors can be signed directly, the clear factors as IDs and bid price, or indirectly, the latent factors evaluated using the utility (CTR) probability model. The considered approaches give the integral feature of a bidding process in the manner of the CLICK probability function. The forecasting function value can be evaluated using its previous samples with account of LR model vector or as a sum of latent factors given by factorization of a matrix which cells contain information about bidding events – interactions between users, advertisers and publishers, events of CLICK and INSTALL. Accumulation of such matrix on condition of big data is difficult process.

The other way is to separate full complex BR stream on independent threads of trades which has stable characteristics on cost and sales. It can be assumed that the products of same shop at same price are characterized by same quality in respect to bidding events probability. So, the model of CLICK and INSTALL prediction may be created for each thread that is indicated as *{shop, product, price}*.

AD consumers can be considered as a members of users classes which are characterized by device. The device is characterized by operating system (OS), its version (OSV), device manufacturer (make) and device model. These parameters form the fields vector: *device={os, osv, make, model}*. Device parameters are important because they effect on AD display. Other information about user may be accounted in a pre-targeting stage by NT schemas. The model of CLICK and INSTALL prediction can be created for each user class.

It is assumed that the combination of effective sources and consumers at a moment of high probability of paid event may give resultative advertising.

Database structure. In mobile application site is signed as a *bundle*. AD banners are signed as creatives with identifiers – *creativeId*. Each site has some banners of different size and visibility. A cost of advertising exhibition on a banner may vary. It is pointed by the start price of the auction and is signed as *minPrice*. So, the vector *{shop, product, price} = {bundle, creativeId, minPrice}* characterizes a source of bidding events.

Table 1. Database thread parameters structure

Field	Comment
rating	rating of click and install
bid, win, click, install	numbers of events
winPrice, winPriceVar	win price and its variation
winbidPrice, winbidPriceVar	win bid price and its variation
lostbidPrice, lostbidPriceVar	lost bid price and its variation
ban	ban counter
wdt=[wdt0,…., wdt23] cdt=[cdt0,…cdt23] idt=[idt0,…idt23]	win, click, install daytime distribution
Plast	last click probability
Tclick	last click time
$[\Delta t_0, \ldots, \Delta t_{N-1}]$	vector of time intervals between clicks
$[\Delta n_0, \ldots, \Delta n_{N-1}]$	vector of bids number between clicks
weight	weight coefficients vector

Each *bundle* parameters can be presented as the cell database. The cell contains some threads signed by *{creativeId, minPrice}*. An example of the cell thread data structure is shown in table 1. The cell array can be stored as structured database. The fragment of the cell includes three threads is presented in Appendix in JSON format.

The consumers database thread is like in table 1 with exclusion of some fields.

Rating filtration. The step-by-step algorithm of bids rating learning and selection of bids with high probability of the events is presented in table 2.

Algorithm starts from learning of registered cell's thread behavioral characteristics. The first operation is reading of data vector pointed by three parameters values from the incoming BR: *{bundle, creativeId, minPrice}*. Each new value of *bundle, creativeId, minPrice* causes creation of new cell or cell's thread in learning mode. The initial *bidPrice* = 1.1·*minPrice* and this price is increasing in the bidding process up to the rating of win to bid reached some level, not less 0.05, for an example. When wins number reached W_{min} value (10 or 20) then a selective filtration can be made by the thread performance evaluation using the specific rating of click with account of installs with their weight and *winPrice*. If *rating* is more than RL_{min} value (0.01 or 0.05) the performance is high and the cell is putting into working mode. Otherwise, the cell is banned for a sometime by the ban time counter. The working mode continue up to *rating* become less then RW_{min} value (0.001 for an example).

The decision that cell's thread is not appropriate for bidding is made on this condition. The system controls the averaged *rating*. If the rating falls, then the *bidPrice* increases in order to increase the number of effective wins. Media buyer can determine the financial limit for participation in trades in the form of a cost of the paid event. Then the constraint *winPrice* < *event_cost* ·*rating* can be defined. This constraint means that maximization of the ratio *max***(**rating/*winPrice***)** is desirable. Each cell thread in table 1 contains three vectors of 24 values: **wdt**; **cdt***;* **idt**. Vectors are designed to fix the distribution of the number of events WIN, CLICK and INSTALL by hours of the day. Instead of total parameters in table 2 points 6, 7 can use values related with current time of bidding, t_b, for example, for *win*: $win(t_b) = wdt[mod_{24}(t_b - 1)] + wdt[t_b] + wdt[mod_{24}(t_b + 1)]$.

Table 2. Algorithm of rating filtration

	Operation	Return	Comment
0	*if* time of cell search expired: new cell creation	*bidPrice* = 1.1·*minPrice*	Bidding is in learning mode
1	*if* thread with parameters *creativeId* or *minPric*e is not found – create new thread		
2	*if ban* > 0 *ban* decrement	"no content"	thread is banned
3	*if ban* = 0	*bidPrice* = 1.1·*minPrice*	bidding is in learning mode
4	*if win* < W_{min}: W_{min} – minimal number of wins for decision making about cell thread performance	if *win/bid* < 0.05 *bidPrice* = 1.1·*bidPrice*	rating increase, W_{min} is 10 or 20
5	*if win* >= W_{min} & *click/win* < RL_{min} *ban=BanCount, bid=win=click=install*=0	"no content"	ban mode of the thread, RL_{min} – minimal rating of ckick in learning mode
	if win >= W_{min} & *click/win* >= RL_{min}	*bidPrice*	thread is in working mode
6	*if win* > W_{min} $rate_t = \frac{click + IW \cdot install}{win \cdot winPrice}$		RW_{min} = 0.02, *IW* – install weight in comparison with click: 200/500/1000
	if $rate_t$ > RW_{min}: RW_{min} – minimal rating in working mode	*bidPrice*	bidding
	else ban=BanCount, bid=win=click=install=0	"no content"	ban mode of the cell thread
7	$rate_{new} = \frac{rate_t + win \cdot rate}{win + 1}$ *if* $rating_{new}$ /*rating* < 0.9	*bidPrice* = 1.1· *bidPrice*	raise bid price if *rating* falls Update *rating*
8	*bid* increment *if* win *win* increment *if* click *click* increment *if* (install) *install* increment		Update events counters

The algorithm in table 2 can be defined as soft mode algorithm which includes procedures of learning and selective filtration. The limit state of the soft algorithm is the hard mode algorithm which includes items 6–8 of table 2. It executes only procedures of the selective filtration when all information about event sources and consumers is known.

Click prediction. Main part of expenses is related with frequent won bids which give rare paid events. The most rating sources and consumers do not give streams of paid events. Their paid events pass with some periodicity. Therefore, it is necessary to estimate the periodicity and to choose bids that follow with a similar periodicity. Then number of paid events of installs and corresponding to them events of click wouldn't be decreased significantly. So, the bids selection should be made from the regard point of CLICK prediction. The click events prediction could be made by evaluation of CLICK probability for current bid request by estimation of time intervals and bids number between consecutive clicks. The vectors of time intervals and the vectors of number of bids between click events in the thread database structure are intended for prediction of click. The number of click events relates with number of WIN which depends on the *bidPrice* level. The rating filter regulates the *bidPrice* level so that there are appropriate flows of WIN and CLICK.

The methods of Bayes-Poisson dynamic model, the regression models basing on the method of Support Vector Machines (SVM) with different kernels, the probabilistic methods, such as Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Cross Validation Model (CVM), etc., can be used for the click events prediction. The methods can be joined into a general schema with weights which reflect their accuracy.

Bayes-Poisson selective filter. The Poisson model of the click events stream can be used for click probability estimation at the current time. The conditional probability of a click event at time of bidding t_b on condition that the previous CLICK occurred at time t_c is the next:

$$P(t_b \mid t_c) = 1 - e^{-\lambda(t_c)\cdot(t_b - t_c)} \tag{1}$$

where $\lambda(t_c)$ is the events intensity at the time t_c. The intensity of the Poisson stream of click events can be evaluated using *N* time intervals Δt_i between consecutive CLICKs as following:

$$\lambda(t_c) = \left(\frac{1}{N} \Sigma i = 0...N-1 \Delta t_i \right)^{-1} \tag{2}$$

The intensity (2) is variable and therefore it should be evaluated for each click. The number of intervals *N* is defined as the order of the filter. The stability and effectivity of the model depends on the order.

The absolute probability of a CLICK at the moment t_b with account the probability $P(t_b \mid t_c)$ and click probability $P(t_b, t_c)$ at the moment t_c is defined by the Bayes formula.

$$P(t_b) = P(t_b \mid t_c) \cdot P(t_b, t_c) \tag{3}$$

where the click probability $P(t_b, t_c) = \frac{N_{click}(t_c)}{N_{win}(t_b)}$, $N_{click}(t_c)$, $N_{win}(t_b)$ – the numbers of events at the specified time. The Bayes-Poisson dynamic model shows that during a short interval after the event the probability of the next one is small and eventually tends to a value $P(t_b, t_c)$ which can change due to a change of the number $N_{win}(t_b)$ of wins.

In the case of consumer CLICK prediction there are several threads of the click events with different intensities, four fields of device parameters, for example. They can be considered as conditionally independent. Then with account (3) the total probability of the event can be represented as

$$P(t_b) \approx \sum_i P_i(t_b \mid t_c^{(i)}) \cdot P_i(t_b, t_c^{(i)}) / \sum_i P_i(t_b, t_c^{(i)}) \tag{4}$$

where the index *i* lists all model parameters values which are presented in the query and are used in the probability model. Expressions (3), (4) are known as the Bayes-Poisson regression (BPR). The probabilities $P_i(t_b \mid t_c^{(i)})$ define the weights with which the conditional probabilities are summed, their magnitudes determine the probabilities of the event at time t_b. This expression has to be normalized in order to know whether a high or low probability level was obtained at the time when the current request was received. Since there are a set of events that have taken place at the same parameters values, it can be estimated the probability $P(t_c)$ of the last of them and use it as conditional "one". The condition of a sufficient level of the current event probability can be set so that it is of the same order that probability of held events.

$$P(t_b) \geq P(t_c) \;\; or \;\; P(t_b) \subset O\big(P(t_c)\big) \tag{5}$$

where $O(\cdot)$ is a value neighbourhood. The application creates the following flag by condition (5) state.

$$flag_{poisson} = \begin{cases} +1, \; (5) \; is \; true; \\ -1, \; (5) \; is \; false; \\ 0, \; (5) \; is't \; defined. \end{cases} \tag{6}$$

The flags (6) are defined for bundle threads and for consumers threads: $flag_{poisson.bundle} \cdot flag_{poisson.cons}$.

Prediction of time interval and bids number between click events.

Regression methods. Let there is the vector of time intervals $\mathbf{v} = [\Delta t_i]_{i=0...N-1}$ between consecutive click events. Then the following time interval between last click and the next click can be found using the forecasting methods. The Python *scikit-learn* library allows to predict next value using a current values vector and a feature matrix by the SVM method with some kernels. The feature matrix is formed as multiplication of the matrices compiled by *2N*-1 time intervals.

$$\mathbf{F} = [F_{i,k}]_{i,k=0...N-1} = \left[\sum\nolimits_{j=0...N-1} \Delta t_{i+j} \cdot \Delta t_{k+j}\right]_{i,k=0...N-1}.$$

The predicted value is the function

$$v = v(\mathbf{F}, \mathbf{v}, kernel) \tag{7}$$

The *kernel* can be "*gaussian*", "*polynom*", "*linear*". It should be chosen the kernel which gives the best quality of the prediction.

Hidden Markov Model (HMM). The elements of the vector (6) can be considered as states of the HMM process of time intervals (or number of bids). The Python library *hmmlearn* implements the algorithm of prediction of the most probable state for each of elements of the vector $\mathbf{v}$. It returns the vector $\mathbf{p} = [p_i]_{i=0...N-1}$ of pointers which point on elements of the vector **v** which are most probable as the next states. If Δt_{N-1} is the last time interval then the next most probable interval is the $\Delta t_{p_{N-1}}$. It is similar for number of bids too.

Gaussian Mixture Model (GMM). The probability distribution function (PDF) of the elements of the vector **v** can be considered as a weighted sum of Gaussian functions with different parameters when PDF is not regular. The Python library *scikit-learn* implements the algorithm of prediction of the most probable value for each of elements of the vector **v** in accordance with estimated PDF. It returns the vector $\mathbf{g} = [g_i]_{i=0...N-1}$ of pointers which point on elements of the vector **v** which are most probable as the next ones, this is $\Delta t_{g_{N-1}}$

Integral statistic schema for events prediction. The Poisson model and regression models (7) with kernels – gaussian, polynomial, linear, give four estimates of the forecast value: Δt_{gaus}, Δt_{poly}, Δt_{line}. The HMM and GMM give two additional estimates – Δt_{hmm}, Δt_{gmm}. There is the problem to define the estimate with highest probability of the click event. It can be formed the vector of predicted time intervals between events as

$$\mathbf{dt} = [\Delta t_{gaus}, \Delta t_{poly}, \Delta t_{line}, \Delta t_{hmm}, \Delta t_{gmm}].$$

The vector **dt** can be considered as states of the random process. The metods HMM and GMM can label the most probable states for each element of the vector **dt** by label vectors **p** and **g**. The element $\Delta t_{\max(i)}$ on which maximal number of the labels point can be chosen as most probable. Also, it is similar for number of bids.

General algorithm of bid selection. The condition

$$t_b \geq t_c + \Delta t_{predict} \approx T_{last} \tag{8}$$

where $\Delta t_{predict}$ is the interval which is predicted by the approaches *A-C* with probability (5) not less than P_{last}, is the condition to select the bid at current time for trading. The flag

$$flag_{time.bundle} = \begin{cases} +1, \ (5)\,\&\,(8) \ is \ true; \\ -1, \ (5)\,\&\,(8) \ is \ false; \\ 0, \ (5)\,\&\,(8) \ is't \ defined. \end{cases} \tag{9}$$

can be defined using conditions (5) and (8).

By the same way the flag $flag_{time.cons}$ is defined for current cell thread of consumer fields.

Analogous prediction of the number of bids between click events gives the flags $flag_{bid.bundle}$ and $flag_{bid.cons}$.

The bids streams of any bundle cell and fields cells are unconditional random processes and therefore statistic of number of bids between click events can serve as feature of the following click event. So, the integral flag can be evaluated using the results of prediction:

$$\begin{aligned} flag = {} & w_{pb} \cdot flag_{poisson.bundle} + w_{pc} \cdot flag_{poisson.cons} + \\ & + w_{tb} \cdot flag_{time.bundle} + w_{tc} \cdot flag_{time.cons} + \\ & w_{bb} \cdot flag_{bid.bundle} + w_{bc} \cdot flag_{bid.cons}. \end{aligned} \tag{10}$$

The weight coefficients $w_{..}$ in (10) are defined as relation between number of true states of flags (6), (9) and number of true events. These coefficients give an advantage to most truthful predictor. With account of the rating filtration and flag (10) value, the condition of bid selection is the following:

$$\begin{aligned} & if \ \ rate_{bundle} \geq R_{\min_{bundle}} \ \ \& \ \ rate_{cons} \geq R_{\min_{cons}} \\ & \& \ \ flag > 0 = true \end{aligned} \tag{11}$$

The bidding will be made on true condition, the value of *bidPrice* is returned by the application in the response message, otherwise the state *"no content"* is returned.

3. Experimental results

The AD trades have different properties in respect to impressions on IOS and Android devices. Therefore, two examples of trades were used for the database learning.

The first one includes protocols of 682171 auctions of AD impression in IOS-devices. Each protocol includes records of messages of bid request, response with *bidPrice* value or "no content" message and bid result with information on occurred events. The bids were selected by NT filtration using only AD categories. The impressions gave 8271 CLICKs and 12 INSTALLs. As the number of INSTALLs is so small it is very difficult to predict CLICK with following INSTALL. Therefore, the same files were used in the tests to investigate the learning process by accumulation of results.

The second one includes protocols of the 9217055 auctions of AD impression in Android-devices. The bids were selected by filtration on categories too. The impressions gave 21495 CLICKs and 232 INSTALLs. There were used some of consecutive tests using these files to investigate the learning process.

Finally, the training process was investigated using protocols of 32458326 auctions, protocols files size is more than 106 Gbyte. The number of defined bundles is 75556 items. The database of cells includes up to 167403 threads which are presented in some JSON files with structure like in Appendix 1. The fields list which characterize consumers contains 10905 threads of statistical data in the case of device parameters using.

Table 3 shows results of the off-line bids selection learning in accordance with general algorithm (11) in the soft and hard modes of rating filtration in the case of IOS devices. There are presented initial data (high line) and selected data (low line). Two cases of *N*=4 and *N*=8 in (2) of the Poisson model (1) and HMM, GMM number of states are considered. There were made 10 cycles of learning and selective filtration in the soft mode. The install price was consecutively reduced step-by-step, averaged values are presented. At the next step followed 10 cycles of hard selection were made. The CTR multiplicity and ITR multiplicity show relative changes of the click and install rating of filtered data in respect to initial ones.

Table 3. Database learning for IOS devices

Number of states	**4**		**8**	
Mode	soft	hard	soft	hard
Learning cycles	0–9	10–19	0–9	10–19
Auctions	6821710			
	2404771	123068	2045975	143552
WIN	933620			
	119728	44573	107659	49826
CLICK	82710			
	14 954	8 573	13 382	9 341
INSTALL	**120**			
	48	**40**	**42**	**32**
CTR	0.0886			
	0.1249	0.1923	0.1243	0.1875
CTR multiplicity	**1.490**	**2.170**	**1.403**	**2.116**
ITR	0.000128			
	0.00040	0.0009	0.00039	0.00064
ITR multiplicity	**3.132**	**7.008**	**3.048**	**5.017**
Click price $	0.12			
	0.07	0.04	0.07	0.04
Install price $	**82.04**			
	21.31	**9.00**	**21.86**	**12.76**

Table 4 shows results of the off-line training in the case of Android devices. It was made one cycle of the learning and selective filtration in the soft mode and one cycle of the hard selection. Note, the doubling of CLICK ratio leads to a fivefold increase of INSTALL ratio. As it can be seen from tables, the results with $N = 4$ in (2) are better than in the case of $N = 8$. This fact reflects nonstationary of the bidding process. Same results were fixed for other values of N. The number of states $N = 4$ is taken as the base.

Table 4. Database learning for Android devices

Number of states	**4**		**8**	
Mode	soft	hard	soft	hard
Learning cycles	0	1	0	1
Auctions	9217055			
	2185100	441903	1742920	439486
WIN	473174			
	91735	32 201	78884	31984
CLICK	21495			
	5796	2862	5007	2928
INSTALL	**232**			
	150	**84**	**111**	**78**
CTR	0.0454			
	0.0632	0.0888	0.0634	0.0915
CTR multiplicity	**1.392**	**1.958**	**1.398**	**2.016**
ITR	0.00049			
	0.00163	0.0026	0.00178	0.00244
ITR multiplicity	**3.337**	**5.324**	**3.647**	**4.977**
Click price $	0.14			
	0.09	0.06	0.09	0.06
Install price $	**12.96**			
	3.57	**2.08**	**4.14**	**2.26**

Table 5. Filter training by all trades

Mode	soft	hard
Auctions	32458326	
	6631253	1677394
WIN	2527980	
	397488	145434
CLICK	112633	
	23737	12202
INSTALL	**405**	
	153	**90**
Win rate	0.0779	
	0.0599	0.0867
CTR	0.0446	
	0.0597	0.0839
CTR multiplicity	**1.3404**	**1.8811**
ITR	0.00016	
	0.00038	0.00061
ITR multiplicity	**2.38**	**3.87**
Click price $	0.11	
	0.07	0.05
Install price $	**29.74**	
	11.04	**7.43**

The learned filter was used in training to select trades across a variety of protocols, the results are shown in table 5. It was not reached double increase of the CTR multiplicity and so averaged INSTALL price was not decreased significantly. The simulation of the total trades had shown that total price has trend to reduce from cycle to cycle. The hard mode allows to obtain appropriate cost of the INSTALL for profitable trade.

Figures 1, 2 show events distribution of initial and selected by the general algorithm bidding events on the plane (*minPrice* × *winPrice*) obtained for all trades. Subfigures 1), 2), 3) show distributions of the events WIN, CLICK and INSTALL.

As it follows from figures, there are two types of the auction start price *minPrice*. The first type is in the range of 1$ per thousand impressions. The second type is in the range of 3–10$ per thousand impressions. The *winPrice* changes in the range from 1$ up to 20$ for the first type. The *winPrice* ~ *minPrice* for the second type of price.

The INSTALL distribution has shown that the main part of these events is shifted relatively to the main part of win and click events towards the increase of the *winPrice*, especially for the first type of *minPrice*. The most effective *winPrice* diapason is 10…16$. The high level of the install events corresponds to low level of the win events. The range of the *winPrice* in the bounds of 3…10$ per thousand impressions is interest too, but it includes least of half of WINs. So, the trades are not effective in term of INSTALL obtaining. The main part of install events for the second type of prices is also in the area of high prices *winPrice*.

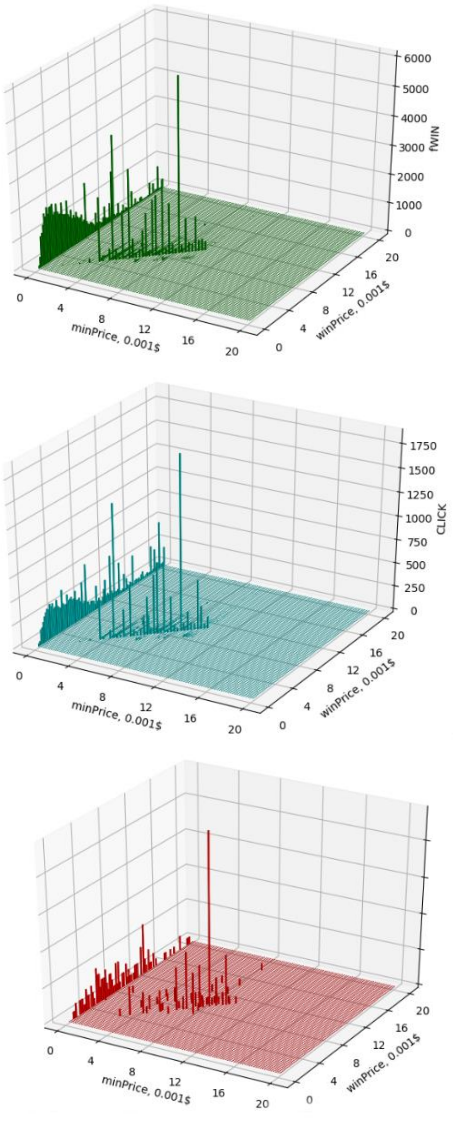


Fig. 1. Price distribution of initial data events

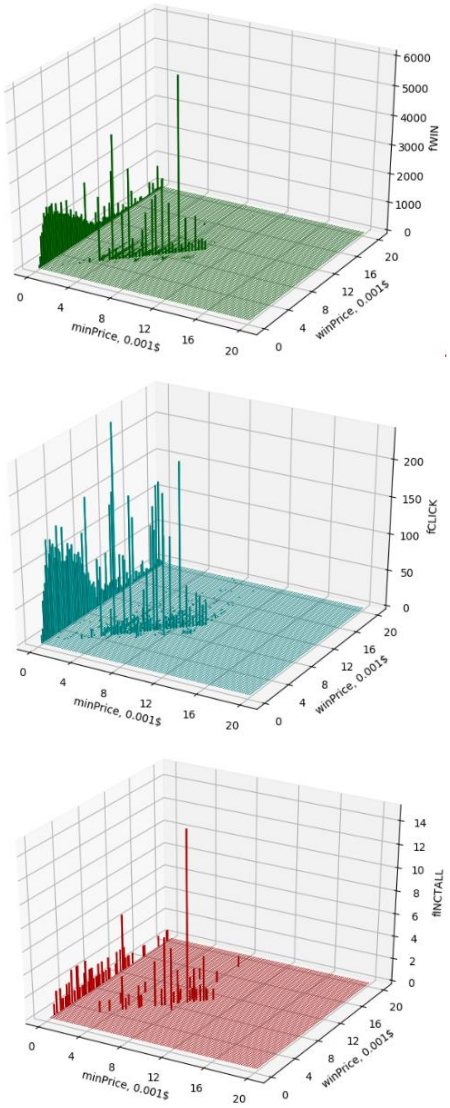


Fig. 2. Price distribution of filtered data events

It can be made the conclusion, that *winPrice* reflects the effectiveness of the trades in terms of obtaining the INSTALL. High level of *winPrice* indicates interest of competitors in trading. So, an estimate of the *winPrice* of current bid is very important characteristic of the bid effectiveness. As statistical analysis has shown, the averaged *winPrice* of each cell thread is relatively stable parameter, for example, its variation is up to 50% when *winPrice* < 4$, 30% when *winPrice* = 4…8$ and less than 10% when price is higher. So, the averaged *winPrice* of previous bids can be used as rating parameter.

The high cost of effective trades highlights the importance of analyzing the likelihood of paid events to trade only when the probability is highest.

As it seen in figures, the most significant peaks of the WIN and CLICK distributions of the selected bids are reduced. This means that intensity of some active sources decreased. These sources are associated with spending a significant part of the funds. The distribution of the INSTALL did not change significantly.

Note. The tracking of the picks in figures 1, 2 has shown, for example, the pick near the left bound on the level of ~15000 wins by the winPrice 13$ per thousand impressions in figure 1.1 corresponds to ~900 clicks in figure 1.2 and 5 installs in figure 1.3. So, the install price is ~39$. The filtration selected the BR which gave ~2500 wins in figure 2.1, ~225 clicks in figure 2.2 and 4 installs in figure 2.3. The install price became ~8$.

4. Conclusion

The main target of the off-line modeling is to determine the dynamic of the bidding process in dependence of system learning. As it follows from the obtained results, the ML model gradually increases the effectiveness of trades in accordance with filling and stabilization of the sources and consumers database.

As statistics has shown, the events cost decreasing is due to the selection of more rating trades and prediction of the events time of occurrence.

The prediction filters play significant role in price decreasing. Mean number of bids between CLICK is up to 300 and it became the same after filtering. But number of WINs decreases up to twice. This allows to reduce the expenses. As statistical investigation has shown, from 70% to 90% of clicks pass through prediction filter with true flags of events.

The hard schema can be used in finish stage of AD session for effectivity improving when all sources are known. For example, when average price of paid events reached some appropriate level. This will mean that the database statistic properties are synchronized with incoming messages property.

Users number which are ready to click and install the proposed product at each moment is bounded. Therefore, it is necessary, at the first, to ensure the fulfillment of the law of large numbers for obtaining a large number of random events of CLICKs and INSTALLs. It is necessary to process about 10–20 millions of bid request messages in each prime-time of one-day auction session, choose among them the most rating ones from the point of view to obtain events and only to make trades on them. The main aim of the statistical analysis is CLICK prediction.

The purpose of targeting advertising is to reduce the price of paid click-install events. The decrease of install price cannot occur at the one-time moment as a result of applying some algorithm of bid selection. The auction bids need to be studied before they will be selected. Therefore, the price of events can be reduced only as a result of the process of studying trades and applying the obtained knowledge to select the most promising advertising bids from the point of view to obtain a CLICK-INSTALL.

Each auction is carried out under certain conditions, which are determined by the composition of competitors, their interests and tactics of bids. Therefore, learning and training should be made at all stages of each session. Historical data should be obtained as averaged over a long period. The transfer of the moment properties of one bidding session to another session may not give positive result.

The first result of the research is that by dividing the total bid request flow into elementary threads as sources of the events and determining their dynamic and general ratings there was created the database in the manner of C++ STL containers or Python lists with supporting access in real time with appropriate delay on decision making. The mean delay time does not exceed 0.5 ms, the maximal time – 1.5 ms correspondingly.

The second result is the classification of AD consumers using parameters of their devices. The statistic characteristics of the parameters values are supported by data containers too.

The third result is application of complex two stage forecasting schema using three parameters and three types of regression based on the SVM method and HMM, GMM for CLICK prediction.

The connection of AD source and consumer with high level of rating in the moment of high level of event probability can give the resultative bidding.

The proposed method implements the AD targeting by the way of paid events prediction. The prediction is basing on evaluation of total and dynamic ratings of paid events along all bidding time with account of historical data and evaluation current event probability in each moment of the bidding. Such approach is usual in professional advertising practice.

The use of a neural network like TensorFlow of a constant structure with overloaded parameters corresponding to the thread of current bid is needed in a long time of decision making on participation in the trades, especially at the learning stage.

References

[1] Adikari S., Dutta K.: Real Time Bidding in Online Digital Advertisement. New Horizons in Design Science 9073, 2015, 19–38.
[2] Avila C. P., Vijaya M. S.. Click Through Rate Prediction for Display Advertisement. International Journal of Computer Applications 136(1), 2016, 18–24.
[3] Bisikalo O., Kharchenko V., Kovtun V., Krak I., Pavlov S.: Parameterization of the Stochastic Model for Evaluating Variable Small Data in the Shannon Entropy Basis. Entropy 2023, 25, 184 [http://doi.org/10.3390/e25020184].
[4] Chapelle O.: Offline Evaluation of Response Prediction in Online Advertising Auctions. IW3C2, Florence, 2015, 943–944.
[5] Chapelle O., Manavoglu E., Rosales R.: Simple and scalable response prediction for display advertising. Transactions on Intelligent Systems and Technology (TIST) 5(4), 2015, Article No. 61, A1–A34.
[6] IAB 2014. OpenRTB API Specification Version 2.2. http://www.iab.net/media/file/
[7] Jahrer M., Töscher A., Lee J.-Y., Deng J., Zhang H., Spoelstra J.: Ensemble of collaborative filtering and feature engineered model for click through rate prediction. Proceedings of KDD Cup 2012 Workshop, Beijing 2012, 1222–1230.
[8] Juan Y., Zhuang Y., Chin W.-S., Lin C.-J.: Field-aware Factorization Machines for CTR Prediction. RecSys'16, Boston, 2016, 43–50.
[9] Kondakindi G., Rana S., Rajkumar A., Ponnekanti S. K., Parakh V.: A Logistic Regression Approach to Ad Click Prediction. Machine Learning Project, 2014, 399–400.
[10] McMahan H. B., Holt G., Sculley D., Young M., Ebner D., Grady J. et. al. Ad Click Prediction: A View from the Trenches. KDD'13, Chicago, 2013, 1222–1230.
[11] Nigam K. L., Afferty J., McCallum A.: Using maximum entropy for text classification. IJCAI-99 1, 1999, 61–67.
[12] Pan Z., Chen E., Liu Q., Xu T., Ma H., Lin H.: Sparse Factorization Machines for Click-through Rate Prediction. IEEE 16th International Conference on Data Mining, 2016, 400–409.
[13] Richardson M., Dominowska E., Ragno R.: Predicting clicks: estimating the click-through rate for new ads. ACM, 2007, 521– 530.
[14] Sree Vani M.: Prediction of Mobile Ad Click Using Supervised Classification Algorithms. International Journal of Computer Science and Information Technologies 7 (2), 2016, 623–625.
[15] Ta A.-P.: Factorization Machines with Follow-The-Regularized-Leader for CTR prediction in Display Advertising. IEEE International Conference on Big Data, 2015, 2889–2891.
[16] The Real-Time Bidding (RTB) Protocol specification, 2016 https://developers.google.com/ad-exchange/rtb
[17] Zhang W., Yuan S., Wang J.: Optimal Real-Time Bidding for Display Advertising. KDD'14, New York, 2014, 1097–1105.

D.Sc. Roman Kvyetnyy
e-mail: rkvetny@sprava.net

Professor of Department of Automation and Intelligent Information Technologies, Vinnytsia National Technical University. Scientific interests include modeling of complex systems and decision-making under conditions of uncertainty (probabilistic and interval methods), modern methods of data processing. The main direction of scientific activity is development of methods and tools for mathematical modeling and information processing in computerized systems of automation and control.

http://orcid.org/0000-0002-9192-9258

Ph.D. Yuriy Bunyak
e-mail: iuriy.buniak@gmail.com

Spilna Sprava Company, Vinnytsya.
The main direction of scientific activity is investigation of the methods of optimization in big data, signals and image processing – denoising, deblurring, object recognition, also using big data.

http://orcid.org/0000-0002-0862-880X

Ph.D. Olga Sofina
e-mail olsofina@gmail.com

Ph.D., senior lecturer of Department of Automation and Intelligent Information Technologies, Vinnytsia National Technical University
The main direction of scientific activity is modern methods of data and image processing, namely methods of filtering textured images and identifying extraneous objects on their background, as well as methods of removing blurring of the image.

http://orcid.org/0000-0003-3774-9819

Ph.D. Oleksandr Kaduk
e-mail: o.kaduk@gmail.com

Associate professor of Computer Engineering Department, Vinnytsia National Technical University,
The main direction of scientific activity is modern methods in development of reliable AC and DC conversions, data and image processing, namely methods of filtering textured images and identifying extraneous objects on their background, as well as methods of removing blurring of the image.

http://orcid.org/0009-0001-2388-9813

Ph.D. Orken Mamyrbayev
e-mail: morkenj@mail.ru

Deputy Deputy General Director in science and Head of the Laboratory of Computer Engineering of Intelligent Systems at the Institute of Information and Computational Technologies of the Kazakh National Technical University named after K. I. Satbayev and associate professor in 2019 at the Institute of Information and Computational Technologies. Main research field: machine learning, deep learning, and speech technologies.

http://orcid.org/0000-0001-8318-3794

M.Sc. Vladyslav Baklaiev
e-mail: vladvlad03072000@gmail.com

Software Engineering, Taras Shevchenko National University of Kyiv, Ukraine
Research area: moderate experience in development of console and web applications. Basic understanding of Spring framework. Able to provide CI/CD of applications based on Maven phases and Docker-Compose. Python: used for implementing academic level scripts, applications and delimiter separated files. JavaScript: Used in combination with framework Vue.js for implementing front-end projects.

http://orcid.org/0009-0008-5767-6964

Ph.D. Bakhyt Yeraliyeva
e-mail: yeraliyevabakhyt81@gmail.com

Senior lecturer of the Information Systems Department, Faculty of Information Technology, M. Kh. Dulaty Taraz Regional University, Taraz, Kazakhstan.
In 2023 completed a full course of Ph.D. program in "Automation, electronics, electrical engineering and space technology" at Lublin University of Technology (Lublin, Poland), and works as a senior lecturer at Taraz Regional University named after M.Kh. Dulati. The direction of the dissertation research was based on the development of a method for measuring spatial strain in composite materials operating under mechanical perturbation.
Research interests: fiber optic technologies, information systems, Internet of Things and blockchain technologies.

http://orcid.org/0000-0002-8680-7694

http://doi.org/10.35784/iapgos.4279 *received: 20.07.2023 | revised: 29.08.2023 | accepted: 27.11.2023 | available online: 20.12.2023*

THE EFFICIENCY AND RELIABILITY OF BACKEND TECHNOLOGIES: EXPRESS, DJANGO, AND SPRING BOOT

Dominik Choma, Kinga Chwaleba, Mariusz Dzieńkowski
Lublin University of Technology, Department of Computer Science, Lublin, Poland

Abstract. *Increasing popularity of web applications has led to the development of many technologies that enable their production, both on the client and server side. This article attempts to compare three most popular server-side frameworks – Django, Spring Boot and Express. Each of the selected technologies is based on a different programming language. These frameworks were compared in terms of request processing time and reliability. Within the conducted research three backend applications handling HTTP requests were created, all of them using the same database consisting of employees' data. Afterwards, a series of load tests was performed to determine levels of efficiency and reliability of created applications for various numbers of virtual users sending requests to the server at the same time. Five test cases with the following number of requests: 1000, 2000, 4000, 8000, and 16000 were planned and performed for each type of HTTP requests handled by the server simultaneously. Based on the obtained results, it was concluded that the Spring Boot framework was the best in terms of request processing time and high reliability. However, it was noted that for many test cases under extreme load, it had a significantly higher percentage of incorrectly processed requests compared to the Express application, even though the application was noticeably slower. The worst results were observed for Django because the test application created for this framework revealed the longest requests processing time and the highest error rate during processing requests out of the three tested applications. The performed studies helped to determine the efficiency and reliability of the tested technologies at various levels of load. Furthermore, the studies were crucial in obtaining knowledge about the evaluated frameworks as well as their properties and formulating conclusions that will be able to help the developers choose technologies before the implementation of their programming projects.*

Keywords: efficiency, reliability, request processing time, Spring Boot, Express, Django

WYDAJNOŚĆ I NIEZAWODNOŚĆ TECHNOLOGII WYTWARZANIA APLIKACJI INTERNETOWYCH STRONY SERWERA: EXPRESS, DJANGO ORAZ SPRING BOOT

Streszczenie. *Wzrastająca popularność aplikacji internetowych doprowadziła do powstania wielu technologii umożliwiających ich wytwarzanie, zarówno po stronie klienta jak i serwera. W niniejszym artykule podjęto się dokonania porównania trzech najbardziej popularnych szkieletów programistycznych strony serwera – Django, Spring Boot, Express. Każda z wybranych technologii opiera się na innym języku programowania. Szkielety zostały porównane pod względem czasu obsługi żądań i niezawodności. W ramach przeprowadzonych badań utworzono trzy serwerowe aplikacje testowe realizujące obsługę żądań HTTP i wykorzystujące tę samą bazę danych, zawierającą dane pracowników. Następnie wykonano serię testów obciążeniowych pozwalających określić wydajność i niezawodność napisanych aplikacji dla różnych liczb wirtualnych użytkowników wysyłających żądania do aplikacji w tym samym momencie. Zaplanowano scenariusze testowe zakładające następujące liczby żądań: 1000, 2000, 4000, 8000 oraz 16000, wykonanych dla każdego z obsługiwanych przez aplikacje testowe typów żądań HTTP. Na podstawie otrzymanych wyników wywnioskowano, że szkielet programistyczny Spring Boot cechuje się najwyższą prędkością wykonywania żądań oraz wysoką niezawodnością. Jednak zauważono także, że dla wielu przypadków testowych przy ekstremalnym obciążeniu miał on wyraźnie wyższy odsetek błędnie obsłużonych żądań w porównaniu z aplikacją utworzoną na bazie szkieletu Express, pomimo że ta była znacznie wolniejsza. Najsłabsze wyniki zaobserwowano dla Django, ponieważ aplikacja testowa opracowana na podstawie tego szkieletu uzyskała zarówno najdłuższe czasy, jak i najwyższy odsetek błędów podczas obsługi żądań spośród wszystkich trzech testowanych aplikacji. Wykonane badania pozwoliły określić wydajność oraz niezawodność przebadanych technologii przy różnych poziomach obciążenia, pozwoliły poznać działanie i właściwości testowanych szkieletów oraz sformułować wnioski, które mogą pomóc deweloperom w doborze technologii przed realizacją ich projektów programistycznych.*

Słowa kluczowe: wydajność, niezawodność, czas obsługi żądań, Spring Boot, Express, Django

Introduction

The widespread availability of the Internet has resulted in the popularity of Internet applications, which facilitate the use of a variety of services directly through a web browser. This eliminates the need to install additional software or consume device hardware resources. Internet applications are software programs hosted on remote servers, and users can access them through a graphical user interface displayed in a browser window. Communication between the browser and web services is facilitated by programming interfaces, commonly referred to as Application Programming Interfaces (APIs). Web applications are dependent on sending requests and receiving responses using various protocols, including the Hypertext Transfer Protocol (HTTP). Over time, numerous programming languages have been developed to facilitate the creation of Internet applications. However, using a pure language would necessitate building all of the necessary functionalities from the ground up. This results in numerous issues and takes up a significant amount of time. Therefore, developers typically employ pre-existing solutions that provide tried-and-tested features that can improve both performance and security.

To improve the efficiency of software development, developers use frameworks that enhance performance of web applications and provide them with greater ease of use. A framework consists of a set of components offering various functions and capabilities to developers. These functions include database management, performing operations on databases, as well as authentication and authorization mechanisms that enhance the security of applications.

A common approach is to divide web applications into two parts: the client-side (frontend) and the server-side (backend). The client-side is responsible for sending requests to the server, receiving and processing responses, and presenting data to the user in an appropriate format. On the other hand, the server-side manages requests, communicates with the database, processes data, and generates responses. This separation allows for the parallel development of both parts of the application, which can reduce the overall software development time. When selecting a framework, it is important to consider the specific requirements of a project, as each of them provides slightly different solutions that can significantly impact parameters such as performance, reliability, maintainability, and portability. These factors contribute to the overall quality of the final product.

The authors of this paper researched the performance and reliability of the most widely used server-side frameworks for JavaScript, Python, and Java. These tests aimed to determine which of the tested frameworks would be the most suitable choice under specific test conditions.

1. Literature overview

Currently web applications have attained a high degree of prevalence, which is associated with the multitude of available server-side tools utilized for their development. The selection of the most suitable technology that satisfies the requirements

of software is one of the primary decisions that must be taken to create a fully functional web application. This process is closely connected with the choice of a suitable programming language and framework. There are numerous scholarly articles addressing the topic of choosing optimal tools for the development of web applications with particular levels of complexity and scope of operations.

A crucial part of the web application development process is the design and implementation of its server-side logic. This subject was considered in the article [5]. The author describes this process in many aspects, including an examination of web applications within the contexts of both static and dynamic websites, the utilization of programming frameworks, and their helpfulness in accessing databases. Then, a more detailed explanation is provided regarding programming frameworks, which includes the characteristics of Express, Spring Boot, and Django and the comparison of their speed in executing a single statement to the database. What is more, the significance of the correct database design is mentioned in the conclusions.

In a subsequent publication, referenced as [2], the authors conducted a comprehensive comparison of four frameworks (Laravel, Ruby On Rails, Django, Spring) based on a three-point scale established by them. The assessment covered various aspects of each framework, including code generators, popularity, business trends, integration with additional software, and plug-in support. According to the authors, Spring received the highest score among the evaluated frameworks, primarily due to its business trends and popularity among programmers. The authors also rate criteria such as scalability, entry threshold for novice programmers, and popularity on specific internet platforms (Haker News, Google Trend, Reddit, GitHub, StackOverflow). In these criteria, Spring ranked third, behind Laravel and Django. The authors concluded that Django was the best framework among those analyzed, citing its ease of use for novice programmers and its adaptability for large web applications.

Articles [1, 4, 6] compare currently popular web application development technologies to identify their advantages and disadvantages. The analysis was carried out based on prepared test applications implementing CRUD (Create, Read, Update, Delete) functionalities. Almost all applications [1, 4] were connected to the database. Only in [6] the database was not used for fear of a possible slowdown in the response and interference in the measurements. The main examined parameter of the performed operations was efficiency. In the case of applications using a database, GET, POST, PUT, and DELETE requests were considered. On the other hand, in the article [5], the bandwidth and its impact on Internet applications and computer resources used were also examined. Article [4] additionally presented the assessment of authentication and authorization, where individual aspects were assessed using a point system. Articles [1, 6] presented the use of JMeter software to simulate virtual users utilizing the created applications.

The comparison of the performance of popular frameworks has been also the subject of scholarly articles [3, 7]. In both publications, two test applications were implemented for the surveyed frameworks, enabling a connection with the database. Furthermore, the Apache JMeter tool was utilized for conducting the tests, as previously mentioned. The first article described the evaluation of the efficiency of the GET, POST, PUT, and DELETE statements for the REST application. In the second article, the efficiency was surveyed using GraphQL as well. The CPU and memory usage was established, and in addition, the first article's tests were conducted for different loads defined by the number of users (1, 8, 64, 128, 256, 512, 1,024). To obtain reliable results, each test was carried out 10 times. The second article presented the parallel execution of requests (100, 250, 500) with varying numbers of rows (1, 50, 100). The tests executed in both articles allowed for the efficiency comparison of frameworks applied for developing the server-side of web applications.

Choosing a technology sufficiently advanced for this application can help in the performance not only for a beginner but also for an experienced programmer. In addition, it is essential to consider the number of elements that are required to combine various programming frameworks and to calculate them for multiple applications. In this study, the authors decided to compile a list of the most popular technologies currently used for web applications on the server side. This was achieved by using test applications on these frameworks and then compiling the results. It is worth noting that this research stands out due to the selection of other tested technologies, their respective versions, and test cases.

2. Aim, hypotheses, scope of work

The aim of this work is to perform a comparative analysis of backend frameworks: Express, Spring Boot, and Django, in their latest stable versions at the time of conducting the research. The efficiency of three test applications, which have been created based on selected backend frameworks, will be compared according to the applied load, depending on the number of requests sent.

The following research hypotheses have been formulated:

1) Express, due to working in the Node ecosystem, which was designed in order to optimize efficiency and scalability of the web applications, is distinguished by better efficiency compared to Spring Boot and Django in the case of a significant number of requests sent to the server.
2) Spring Boot, by utilizing configuration based on annotations, is characterized by the best efficiency compared to Express and Django in the case of a limited number of requests.
3) Django, due to the significant usage of network bandwidth, is marked by the worst efficiency compared to Express and Spring Boot regardless of the number of requests sent.

3. Used technologies and tools

To compare the chosen programming frameworks, identical applications have been developed using technologies such as Spring, Express, and Django based on a REST architectural style. The established test applications consist of the same functionalities. A tool Apache JMeter has been utilized to simulate requests sent from the client-side to the server applications and to measure response time to these requests.

3.1. REST

REST [13] (REpresentational State Transfer) is an architectural style that defines the way web applications are created to be smoothly usable and user-friendly. REST is an implementation of this architecture that uses HTTP protocol to perform operations on resources and returns data in formats such as JSON or XML, enabling their seamless utilization by client-side applications.

3.2. Apache JMeter

JMeter [8] is an open source software specifically designed for testing the efficiency of an application. It enables conducting tests that simulate user traffic and measure the efficiency of a tested system under load. Test cases and parameters can be adjusted based on the user's preferences. What is more, JMeter provides various tools for analyzing the obtained test results such as generated charts or reports. It is stated to be a popular choice for testing multiple protocols, including HTTP. Moreover, it could simulate virtual users who use the tested system. Weaknesses of the system can be identified, which is a crucial step in taking measures to improve the application’s performance.

3.3. Compared frameworks

After analyzing the popularity of the available frameworks on the market it can be observed that Express, Spring Boot, and Django belong to the top-tier technologies used in backend application development [11]. The comparison was conducted by

examining stars given to the official repositories of the surveyed frameworks by GitHub users [10], and votes received in Stack Overflow's annual summary of the programming market for 2022 [12]. This comparison is presented in table 1. It can be noticed that the discussed frameworks are similarly popular. Django is the most popular framework based on data obtained from GitHub, while Express is the most recognized according to the Stack Overflow survey.

Table 1. A comparison of the popularity of the surveyed frameworks based on data obtained from GitHub and Stack Overflow

Framework/Service	GitHub (stars)	Stack Overflow (received votes)
Express	59,600	22.99 %
Spring Boot	65,100	16.13%
Django	68,200	14.65 %

4. Research methodology

Created applications built on the REST architecture were deployed in a designated testing environment. They implement the basic CRUD methods using the same database. The implemented functionalities were tested with varying numbers of user requests sent within a one-second interval, in order to examine the correlation between the number of requests made and response time.

4.1. Test environment

The research was conducted using a computer with Windows 10 operating system installed. Table 2 presents the parameters that are crucial from the perspective of the performed research. On the other hand, Table 3 provides information about the versions of the frameworks used in the research, which were the latest stable versions available at the time, along with their corresponding programming languages.

Table 2. Parameters of the computer utilized in the research

Parameters	Device
Processor	Intel Core i5-10210U
RAM memory	16 GB
Operating system	Windows 10

Table 3. A characteristic of the tested frameworks

Framework	Version	Language	Version
Express	4.18.2	JavaScript	ES6
Spring Boot	3.0.2	Java	17
Django	4.1.6	Python	3.11

4.2. Test cases

A comparative analysis of the backend application development technologies - Express, Spring Boot, and Django – was conducted using the Apache JMeter tool. The test cases were planned based on common HTTP requests sent by the client to the server. A comparison was performed for the following test cases:

1) the measurement of the execution time of a GET request,
2) the measurement of the execution time of a POST request,
3) the measurement of the execution time of a PUT request,
4) the measurement of the execution time of a DELETE request.

The goal of that research is to examine how the application works in variable testing conditions - using various loads. According to the analysis of the literature, it was decided to undertake test cases for the following cases:

1) sending 1 request by 1,000 users at the same time,
2) sending 1 request by 2,000 users at the same time,
3) sending 1 request by 4,000 users at the same time,
4) sending 1 request by 8,000 users at the same time,
5) sending 1 request by 16,000 users at the same time.

Apart from efficiency, the reliability of an application is known to be a crucial characteristic. It was also the subject of the research in this study.

4.3. Test applications

The created test applications were based on a fragment of an open-source MySQL database called Employees [9], which consists of approximately 4 million records. The diagram representing a part of the database containing information about employees and their salaries is presented in figure 1. The employees table contains employee data, and the salary table is related to it through a one-to-many relation which comprises information about employee salary.

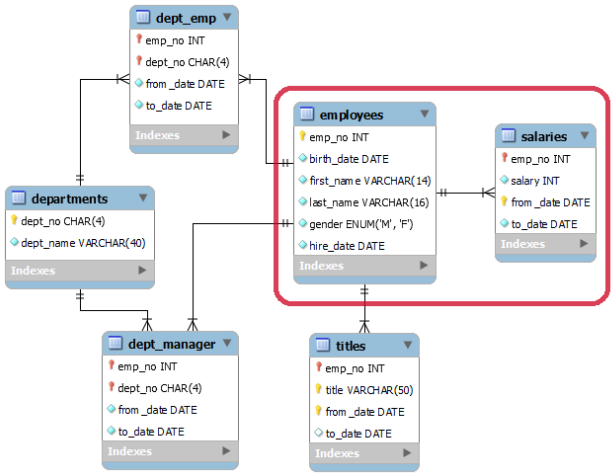


Fig. 1. The diagram presenting a database with a selected fragment, which the application utilizes

The GET request implemented in the study retrieves employee data using the identification number specified in the path. The request returns all details pertaining to the selected employee, along with a list comprising information on their remuneration. Conversely, the POST request is responsible for adding a new employee, achieved by sending an appropriate JSON object containing the user's data in the request's body. The PUT request, on the other hand, modifies the employee data associated with the employee ID specified in the path, replacing it with the data provided in the sent JSON object. Finally, the DELETE request deletes the employee's data with the indicated ID, along with all information concerning their earnings.

During the application testing, it was observed that certain development frameworks implement mechanisms that boost application performance by default, with Spring Boot featuring the largest number of such solutions. Conversely, Express lacks pre-implemented mechanisms of this nature, and their use requires additional programming and configuration efforts. To enhance the efficiency of the test application based on Express, a clustering mechanism was developed, allowing for the launch of multiple application instances within a single process. This mechanism improved overall performance while simultaneously reducing resource consumption.

5. Results analysis

The conducted load tests of the three implemented applications allowed for an analysis of the surveyed frameworks in terms of both the speed of request execution and determining their reliability.

5.1. GET request

The results for the GET request are pictured in figure 2, with a horizontal axis representing the test load and a vertical axis representing the average request time.

In this case for each of the conducted test cases, the application written using Spring Boot was characterized by the best request execution time. The Express framework application was distinguished by its worse efficiency, while the request execution times for the subsequent test cases are

significantly longer compared to those of the applications developed using Spring Boot. The least favorable results in terms of efficiency were obtained from the Django application. The operation execution times are approximately twice the times achieved for the application implemented in Express. Despite the number of requests, the average request execution time ranged between 2,800 and 4,100 ms.

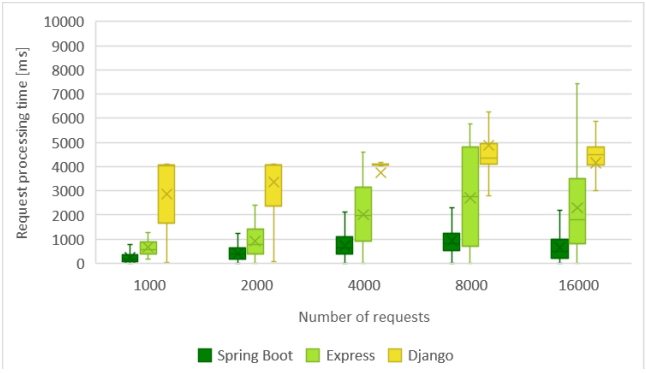


Fig. 2. The GET request processing time depending on the number of requests for each of the tested frameworks

Figure 3 depicts the reliability tested for the GET request, with the vertical axis representing the percentage of wrong requests.

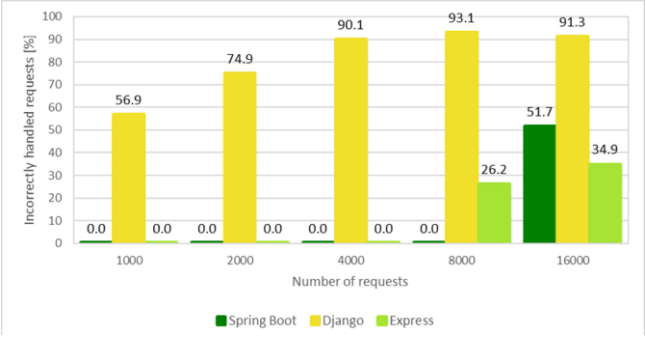


Fig. 3. The percentage of incorrectly handled GET requests by the application based on a given framework

The test application developed using the Spring Boot framework exhibited a high level of reliability, as errors only appeared in the last test case (Fig. 3). For the Express frameworks, errors were noticed for 8,000 users sending requests simultaneously, with approximately 26.23% of requests failing. However, for the server loaded with requests from 16,000 users, the number of unsuccessful requests was lower for this framework than for the Spring Boot framework (34.86% for Express and 51.65% for Spring Boot). On the other hand, the application created with Django was identified by the error occurring at the stage where fewer virtual users were simulated compared to the previous two applications.

5.2. POST request

The results obtained for the GET and POST request were similar. For the evaluated handling of the POST request, the outputs were presented in Figure 4, where the horizontal axis represents the test load, and the vertical axis shows the average time of one request.

It has been acknowledged that the test application based on the Spring Boot framework exhibited the best request execution time. On the other hand, the application utilizing Express demonstrated slightly worse reliability. Although the outcomes marginally varied for an insignificant number of requests sent simultaneously (1,000-2,000), the discrepancy considerably increased for a greater number of virtual users. Among the three test applications, the one based on the Django framework achieved the least favorable result. Execution times for all test cases were markedly higher and substantially deviated from the outputs acquired for the remaining two applications – the obtained times ranged between 3,000-5,000 ms.

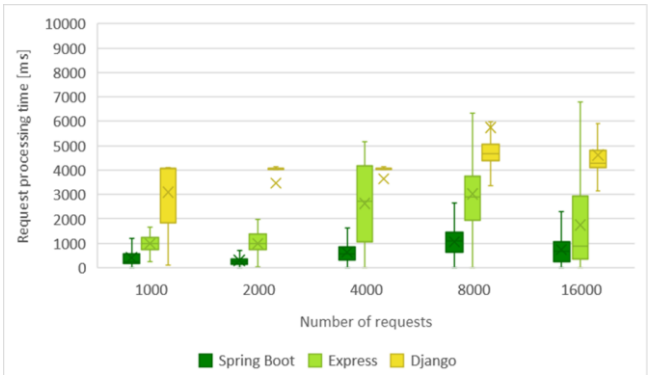


Fig. 4. The POST request processing time depending on the number of requests for each of the tested frameworks

The tested reliability of the POST request was demonstrated in figure 5, with the vertical axis containing information on the percentage of the incorrect requests.

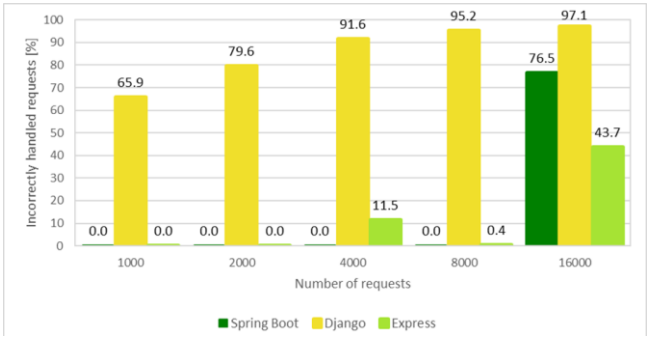


Fig. 5. The percentage of incorrectly handled POST requests by the application based on a given framework

In the conducted tests, the test application utilizing the Django framework exhibited the highest number of failed requests. Specifically, the initial test resulted in a high error rate of 65.9%. Moreover, subsequent tests showed a further increase in the error rate, reaching 97% with 16,000 virtual users. In contrast, the Spring Boot framework exhibited no errors when the server was loaded with 1,000, 2,000, 4,000, and 8,000 virtual users sending requests. Correspondingly, the Express framework demonstrated a zero-error rate for 1,000 and 2,000 requests sent concurrently. An abrupt increase in the number of failed requests was observed during the load simulation involving 16,000 users, as 76% of requests failed for the Spring Boot framework test case. Regarding reliability, Express maintained a zero percent error rate, with a small deviation (11.45% of unsuccessful requests) detected for the test case handling 4,000 requests. Nevertheless, this value decreased to 0.38% for 8,000 requests. For the highest load, 43.69% of requests, sent by 16,000 users at once, were found to be ineffective. Notably, this error rate was lower than that noticed for the application based on the Spring Boot framework under the same load.

5.3. PUT request

Figure 6 shows the average request handling time depending on the number of PUT requests sent. The horizontal axis represents the load levels subjected to experimentation, while the vertical axis corresponds to the average time duration of a single request.

In this instance, it is evident that the Spring Boot framework delivered the most favorable performance outcomes. Nevertheless, it is worth mentioning that when subjected to a server currently handling requests from a thousand virtual users, the Express framework exhibited a noticeably superior processing speed. For other load levels, the average duration of request execution remained stable within the range of 800 to 2,000 ms. Conversely,

Django demonstrated considerably longer request handling times (with the exception of a scenario involving simultaneous requests from 8,000 users), spanning from 3,500 to 4,500 ms.

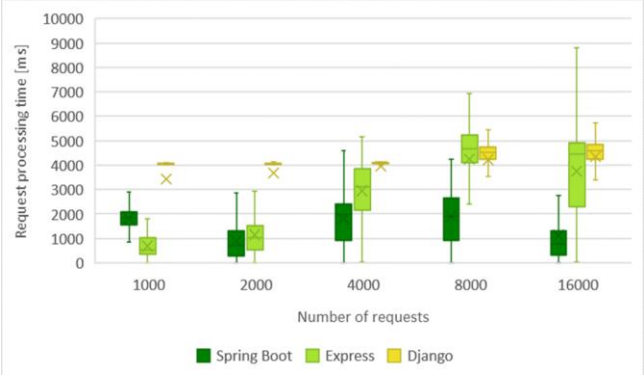


Fig. 6. The PUT request processing time depending on the number of requests for each of the tested frameworks

Figure 6 depicts the reliability data acquired for the PUT request, with the vertical axis displaying information about the percentage of requests that failed. The horizontal axis denotes the number of virtual users whose requests were handled by the server.

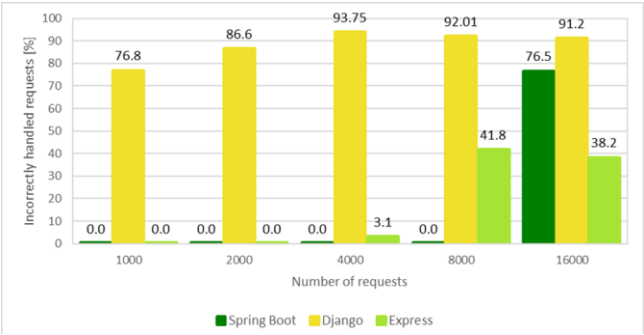


Fig. 7. The percentage of incorrectly handled PUT requests by the application based on a given framework

During the testing, the Spring Boot framework recorded zero failed requests. However, the error rate increased significantly when the server was loaded with 16,000 simultaneous requesting users, with a rate of over 75%, which was considerably higher than that of the Express framework. The Express framework consistently demonstrated average performance with respect to average error rate scores. The only exception was in a case involving 1,000 users requesting a server, where the Express framework outperformed the others. On the other hand, when the server was loaded with 8,000 users, the Express framework exhibited the highest error rate, which was slightly different from that obtained with the Django framework. In the initial three studies, the Express framework exhibited a minimal number of failed requests. However, when the server application was loaded with requests sent by 8,000 and 1,600 users simultaneously, the rate increased to 40%. Conversely, the Django framework fared the poorest of the three evaluated technologies, with a range of failed requests varying between 75% and 90%.

5.4. DELETE request

As a part of the tests for DELETE method requests, the functionality of cascade deletion for specific employees and their corresponding salaries was implemented. The results of this request are illustrated in Figure 8. The vertical axis represents the mean duration of a single request, while the horizontal axis denotes the applied load.

In contrast to prior studies, the results of this experiment did not reveal any distinct disparities in the mean request duration, relative to the used framework and the number of users. Among the three frameworks tested, Express demonstrated the fastest average request execution time of approximately 1,000 ms, for 1,000 virtual users. However, Django and Spring Boot were comparatively slower, with Django exhibiting a difference of less than 500 ms and Spring Boot exhibiting a variance of 1,000 ms. It is noteworthy that Express exhibited the most pronounced decrease in performance among the development frameworks tested, as the number of users increased, ultimately yielding the least favorable results under heavy loads. In contrast, the outcomes obtained with Spring Boot and Django were comparable, with execution time differences for individual tests falling within the range of 1,500 ms. Nonetheless, when tested with 16,000 virtual users, both frameworks exhibited an almost identical mean request execution time, of approximately 3,700 ms.

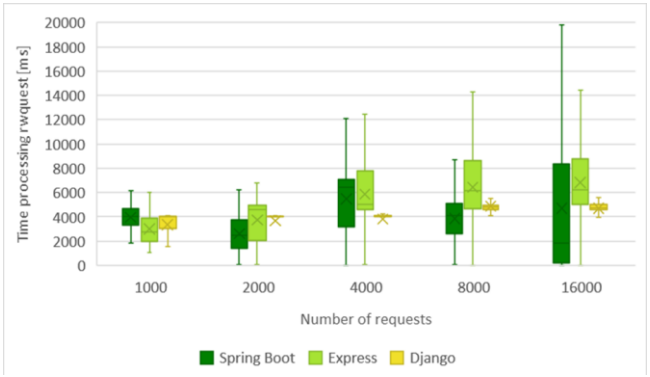


Fig. 8. The DELETE request processing time depending on the number of requests for each of the tested frameworks

Figure 9 displays the data pertaining to the reliability of a PUT request. The vertical axis denotes the proportion of requests that failed, while the horizontal axis represents the applied load under examination.

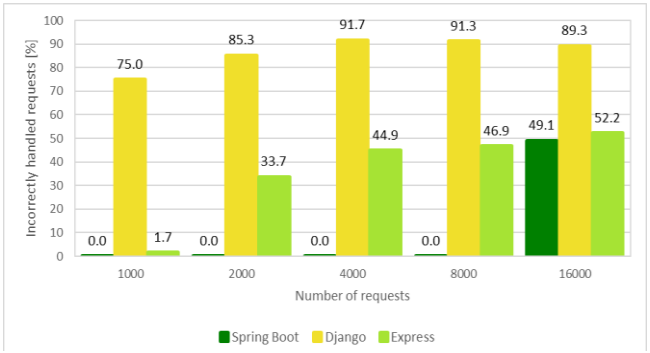


Fig. 9. The percentage of incorrectly handled DELETE requests by the application based on a given framework

With regard to the reliability of the DELETE request, Spring Boot (as depicted in Figure 9) produced the most favorable results, with no errors recorded across request loads ranging from 1,000 to 8,000 virtual users. However, when the server was subjected to a load of 16,000 users, the number of requests that were incorrectly handled amounted to approximately 50%. Notably, Django yielded considerably high error percentages, spanning from the lowest to the highest load tested. Meanwhile, Express exhibited an increasing trend of errors, rising from 0% for the 1,000-user server load to over 50% for a load of 16,000 users.

6. Conclusions

As a part of the experiment, three identical test applications were developed, each implementing a connection with the database and handling HTTP requests. Prior to implementing the test applications, a comprehensive literature review was conducted which helped define the type and parameters of these applications. The chosen database features a compact structure containing six tables, which have been populated with a significant amount of data. The selection of technology was predicated on the identification of the most prevalent server-

side programming languages, followed by choosing the top programming frameworks available for them. The conducted surveys made it possible to determine the efficiency and the reliability of tested technologies at various levels of load, obtain knowledge of the operation and properties of the evaluated frameworks, and draw conclusions that can help developers in selecting appropriate technologies for their programming projects.

Based on the obtained results, the following conclusions were formulated:

1) When comparing test cases under extreme load (with the server handling requests from 16,000 virtual users simultaneously), it was observed that Express had a significantly lower rate of failed requests than Spring Boot, although the application developed with the Express framework was noticeably slower.
2) Out of the surveyed frameworks, Spring Boot is identified by the highest request processing speed and high reliability for a server loaded with requests sent by 1,000-8,000 users.
3) The application utilizing the Django framework demonstrated the longest response time and the highest rate of errors during request handling.

Based on the above conclusions, it can be inferred that the research hypotheses have been verified. The superior results obtained with Spring Boot arise from the implementation of performance-enhancing mechanisms from the Spring framework. Unlike the Express framework, these mechanisms are an integrated part of the Spring and are configured and utilized by default. Express, on the other hand, according to its creator's arrangements, is a minimalist framework, which means that by default there are no these types of mechanisms. They are feasible to use thanks to the many libraries available for those purposes installed through the package manager. Nevertheless, that requires a programmer's knowledge of these mechanisms, their selection, and their configuration. Django exhibited the poorest performance compared to the other two technologies. This may be caused by its threaded architecture, which assigns a distinct thread to handle each request. In situations involving substantial loads, this approach results in augmented memory consumption and required processing time.

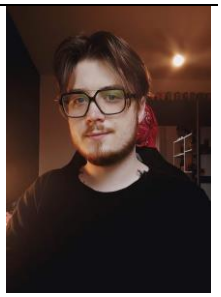
References

[1] Dhalla H. K.: A Performance Comparison of RESTful Applications Implemented in Spring Boot Java and MS.NET Core. Journal of Physics: Conference Series 1933, 2020.
[2] Kaluža M., Kalanj M., Vukelić B.: A comparison of Back-End Frameworks for Web Application development. Zbornik Veleučilišta u Rijeci 7, 2019, 317–332.
[3] Karlsson P.: A performance comparison Between ASP.NET Core and Express.js for creating Web APIs. Jönköping University 2021.
[4] Kopyl P., Rozaliuk T., Smołka J.: Comparison of ASP.NET Core and Spring Boot ecosystems. Journal of Computer Sciences Institute 22, 2022, 40–45.
[5] Muittari J.: Modern Web Back-End. What happens in the back end of the application? Oulu University of Applied Sciences 2022.
[6] Qvarnström E., Jonsson M.: A performance comparison on REST-APIs in Express.js, Flask and ASP.NET Core. Mälardalen University, 2022.
[7] Söderlund S.: Performance of REST applications: Performance of REST applications in four different frameworks. Linnaeus University 2017.
[8] Apache JMeter [https://jmeter.apache.org/] (available: 2023.03.04).
[9] Employees Sample Database [https://dev.mysql.com/doc/employee/en/] (available: 2023-04-18).
[10] GitHub Framework [https://github.com/topics/framework] (available: 2023.01.18).
[11] Most Popular Backend Frameworks – 2012/2022 [https://statisticsanddata.org/data/most-popular-backend-frameworks-2012-2022/] (available: 2022.11.22).
[12] Stack Overflow 2022 Developer Survey [https://survey.stackoverflow.co/2022/] (available: 2023.01.18).
[13] What is REST [https://restfulapi.net/] (available: 2023.01.18).

M.Sc. Dominik Choma
e-mail: dominik.choma@pollub.edu.pl

Dominik Choma received his master's degree in computer science in the area of web application at the Faculty of Electrical Engineering and Computer Science at the Lublin University of Technology. The author's research interests include graphic design, frontend, and backend technologies.

http://orcid.org/0009-0004-6302-5683



M.Sc. Kinga Chwaleba
e-mail: kinga.chwaleba@pollub.edu.pl

Kinga Chwaleba received a master's degree with distinction in Computer Science with specialization in Web Applications at the Faculty of Electrical Engineering and Computer Science at the Lublin University of Technology. The author's research interests include Java, Spring Boot, and backend technologies.

http://orcid.org/0009-0007-3458-5464

Ph.D. Mariusz Dzieńkowski
e-mail: m.dzienkowski@pollub.pl

Assistant professor in the Computer Science Department at the Faculty of Electrical Engineering and Computer Science at Lublin University of Technology. His scientific interests include human-computer interaction, eye tracking applications and web application development. He is a member of the Polish Information Processing Society.

http://orcid.org/0000-0002-1932-297X

http://doi.org/10.35784/iapgos.4421 received: 19.08.2023 | revised: 19.09.2023 | accepted: 27.11.2023 | available online: 20.12.2023

CLOUD TECHNOLOGIES IN EDUCATION: THE BIBLIOGRAPHIC REVIEW

Artem Yurchenko[1], Anzhela Rozumenko[2], Anatolii Rozumenko[2], Roman Momot[1], Olena Semenikhina[1]
[1]Sumy State Pedagogical University named after A.S. Makarenko, Department of Computer Science, Sumy, Ukraine, [2]Sumy National Agrarian University, Department of Higher Mathematics, Sumy, Ukraine

Abstract. *The paper considers the use of cloud technologies in education through the prism of bibliographic analysis. The article characterizes the current state of cloud technologies in education, summarizes the trends, and forecasts the directions of recent scientific research. The leading research methods were bibliographic (visual and quantitative) analysis of keyword networks and qualitative discussion. The bibliographic analysis is based on publications indexed by the scientometric database Web Of Science over the past 20 years. The sample for analysis was formed by searching for the words cloud technology, education, learning, and teaching. The results of the study showed: a significant increase in the popularity of cloud technologies in education in recent years; an increase in the number of studies related to various aspects of educational activities under the influence of Industry 4.0; a gradual increase in the number of studies on the virtualization of the educational process and the use of artificial intelligence in education; dissemination of research on the effectiveness of various types of training using cloud services and teaching methods based on artificial intelligence; the relevance of the trend of visualization of educational material and visual analysis in education. The qualitative discussion provided grounds to identify general trends regarding future research directions.: development of mass online courses and learning technologies (immersive, the use of virtual, augmented, and mixed reality, gaming learning technologies, BYOD approach); further virtualization of universities; development of inclusive education, educational analytics, and assessment (formative and adaptive computer assessment); early training of teachers to use cloud technologies and specialized services in subject learning; research related to visualization (big data, design, simulation, simulation of various processes, etc.) and the designing of relevant new academic disciplines; research of STEM and STEAM education.*

Keywords: cloud technologies, education, learning, teaching

TECHNOLOGIE CHMUROWE W EDUKACJI: PRZEGLĄD BIBLIOGRAFICZNY

Abstrakt. *W artykule omówiono wykorzystanie technologii chmurowych w edukacji pod kątem analizy bibliograficznej. W artykule scharakteryzowano obecny stan wykorzystania technologii chmurowych w edukacji, podsumowano trendy i prognozy kierunków odpowiednich badań naukowych. Wiodącymi metodami badawczymi były wizualna i ilościowa analiza sieci słów kluczowych oraz dyskusja jakościowa. Analiza bibliograficzna została przeprowadzona na publikacjach indeksowanych przez scjentometryczną bazę Web Of Science przez ostatnie 20 lat. Próbkę do analizy tworzy się poprzez wyszukiwanie słów words cloud technology, education, learning, teaching. Wyniki analizy wykazały: znaczący wzrost popularności technologii chmurowych w edukacji w ostatnich latach; wzrost liczby badań związanych z różnymi aspektami działalności edukacyjnej pod wpływem Przemysłu 4.0; stopniowy wzrost liczby badań nad wirtualizacją procesu edukacyjnego i wykorzystaniem sztucznej inteligencji w edukacji; upowszechnianie badań nad efektywnością różnego rodzaju szkoleń z wykorzystaniem usług chmurowych oraz metod nauczania opartych na sztucznej inteligencji; znaczenie trendu wizualizacji materiałów edukacyjnych i analizy wizualnej w edukacji. Dyskusja jakościowa dała podstawy do przewidywania kierunków odpowiednich badań: rozwoju masowych kursów online i technologii uczenia się (immersyjne, wykorzystanie rzeczywistości wirtualnej, rozszerzonej i mieszanej, technologie uczenia się w grach, podejście BYOD); dalsza wirtualizacja uczelni; rozwój uczenia się włączającego, analityki edukacyjnej i oceny (formatywna i adaptacyjna ocena komputerowa); proaktywne szkolenie nauczycieli w zakresie korzystania z technologii chmurowych i specjalistycznych usług w zakresie uczenia się przedmiotów; badania związane z wizualizacją (big data, projektowanie, symulacja, symulacja różnych procesów itp.) oraz rozwój nowych dyscyplin akademickich do prezentacji różnych danych; badanie edukacji STEM i STEAM.*

Słowa kluczowe: technologie chmurowe, edukacja, uczenie się, nauczanie

Introduction

The use of information technology (IT) in education today is commonplace. Scientists explore the features of the use of specialized software in teaching individual disciplines, simplify the organization of the educational process through IT, and automate the control of academic progress. The impact of IT on education has become much more noticeable due to the pandemic. The educational industry was forced to respond due to restrictions on direct communication under external circumstances. School teachers and university professors have begun to master cloud technologies, which today have become the leading tool for providing educational services. The pedagogical literature actively discusses the problem of introducing cloud technologies in learning [32], analyzed the implementation of the BYOD approach [27], analyzed the solution of social, pedagogical, and technical problems that may arise during the implementation of distance [25], mobile and blended learning [42], discusses the feasibility of using cloud data storage (for example, MS OneDrive) [22], experience in using IT to knowledge control (for example, Plickers as a mobile application for reading QR codes [11]). Subject cloud services are studied separately, in particular, such as Geogebra (http://www.geogebra.net) [29], PhET (https://phet.colorado.edu/uk/) [9], Open Source Physics (https://www.compadre.org/osp) [43], Wolfram Demonstrations Project (https://demonstrations.wolfram.com) [38], Virtual Lab (http://chemcollective.org/) [30] etc. These results are accumulated in a significant number and systematized on the use of IT in education in general, which is confirmed by studies: [1] – for the academic environment, [40] – for high school, [34] – for preschool education, [33] – for secondary education and teachers, [20] – for intellectual academic environments, [21] – for the educational sector as a whole in the transition to education 4.0. At the same time, the scientific results of using cloud technologies in education as a type of information technology are not generalized and therefore require a comprehensive analysis.

The article aims to characterize the current state of existing research on the use of cloud technologies in education and identify general trends in future research.

1. Material and methods

The bibliographic (quantitative and visual) analysis of keyword networks and qualitative discussion of the results were used to characterize the landscape of cloud technology use in education. The bibliographic analyses were carried out through the VOSviewer application (https://www.cwts.nl). This app visualizes bibliographic data and builds the set of keywords and their relationships.

The type of analysis – Co-occurrence. The unit of analysis – All keywords.

We took publications from the Web of Science database for analysis. The publications were dated 2004-2023 and related to the use of cloud technologies in education. Keywords used to search for publications in the database Web Of Science: cloud technology, education, learning, teaching. The search date in the Web of Science database is July 5, 2023.

The conducted research is limited to publications presented in Web Of Science over the past 20 years. It does not include articles from other databases, as well as articles written not

in English. At the same time, the analysis expresses the trends in using cloud technologies in education. It gives an idea of the impact of cloud technologies on the educational industry as a whole and the possibilities of using cloud technologies for learning and teaching.

2. Results of research

We realized the search the database for the words "education, modern education, contemporary education, current education", which could be found in the annotations to publications. In total, 1279 results were found. After 2016, there were more than 100 of them annually. The total number of keywords is 4312. If you make a limit of 10 to the depth, then only 59 keywords form a network, and each word occurs at least ten times (Fig. 1).

The network has several clusters: green – education (history, skills, critical thinking, training); yellow – technology (teacher, learning, innovation); violet – simulations (model, control); light blue – digitalization (creativity, school, influence, educational policy); red – higher education (distance learning, educational systems, educational reforms); blue – curriculum (student, teacher training, inclusion, inclusive education, quality).

The network analysis shows a strong connection between "education" with the keywords "technology, teacher, distance learning, and higher education".

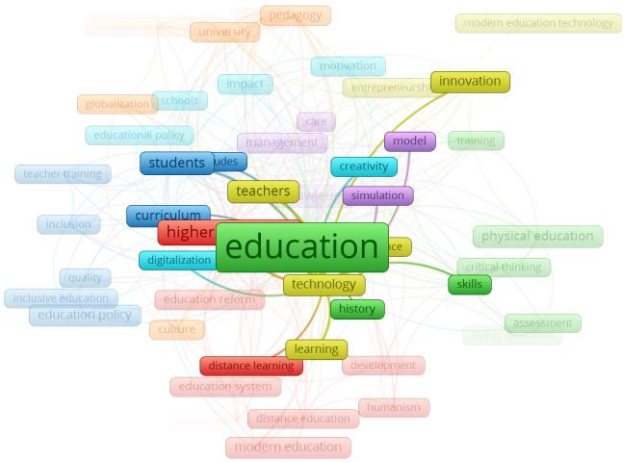
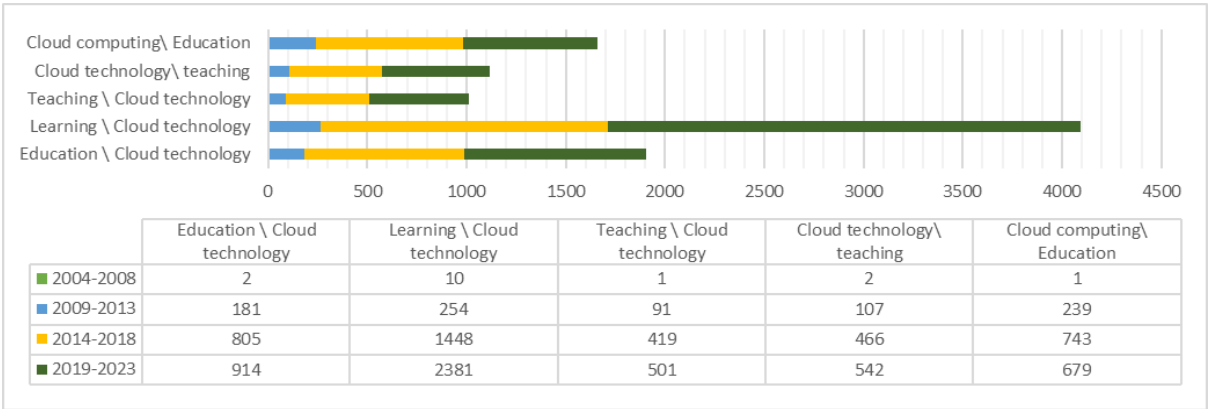


Fig. 1. Network by searching "education, modern education, contemporary education, current education"

Table 1. Quantitative search characteristics

No	Topic\ Abstract	1*	2*	3*	4*	5*	Main clusters	
1	Education \ Cloud technology	1870/5380	20	49	**A**	5	*Violet cluster*: *Green cluster*: *Red cluster*:	cloud services, IT, innovations cloud computing, mobile learning, education technology, virtualization, online, mobile learning internet, big data, artificial intelligence, system, management, augmented reality
2	Learning \ Cloud technology	4093/12214	40	50	**B**	5	*Violet cluster*: *Green cluster*: *Blue cluster*: *Red cluster*: *Yellow cluster*:	Industry 4.0 security, optimization, edge computing machine learning, IoT, big data higher education, education technology, online, motivation Internet, IoT, technologies, 5G, cloud
3	Teaching \ Cloud technology	1012/2955	10	43	**C**	6	*Violet cluster*: *Green cluster*: *Blue cluster*: *Red cluster*: *Yellow cluster*:	cloud technology, mobile learning, distance learning, virtual reality artificial intelligence, Industry 4.0, IoT, big data cloud computing, virtualization, flipped classroom education, education technology, e-learning, performance, teaching, students, higher education technology, innovation, pedagogy
4	Generally	6975/14765	40	73	**D**	3	*Red cluster*: *Green cluster*: *Blue cluster*:	cloud computing, education, information technology, management, simulation, design, mobile learning, Industry 4.0, virtualization internet, IoT, cloud, fog computing, security machine learning, algorithm, system, big data
5	Cloud technology\ Teaching	1117/3332	10	60	**E**	5	*Red cluster*: *Green cluster*: *Blue cluster*: *Yellow cluster*: *Violet cluster*:	cloud computing, education, information technology, management, simulation, design, mobile learning, Industry 4.0, virtualization, flipped classroom internet, big data, augmented reality, artificial intelligence, IoT, gamification, university system, cloud technology, management, design, model, systems technology, higher education, e-learning, online
6	Cloud Computing \ Education	1662/4603	10	105	**F**	7	*Yellow cluster*: *Red cluster*: *Green cluster*: *Blue cluster*: *Light blue cluster*:	cloud computing, technology, higher education, PAAS, SAAS, computer science, education cloud, virtualization, distance learning, information technology adoption, cloud computing adoption, extension, social media, behavioral intention internet, IoT, machine learning, artificial intelligence, virtual reality, learning big data, implementation technology, higher education

1* Number of search results\ Number of keywords; 2* Network depth (number of keyword repetitions); 3* Number of network keywords; 4* Network of connections; 5* Number of clusters in total



	Education \ Cloud technology	Learning \ Cloud technology	Teaching \ Cloud technology	Cloud technology\ teaching	Cloud computing\ Education
2004-2008	2	10	1	2	1
2009-2013	181	254	91	107	239
2014-2018	805	1448	419	466	743
2019-2023	914	2381	501	542	679

Fig. 2. Quantitative data on publications based on search results

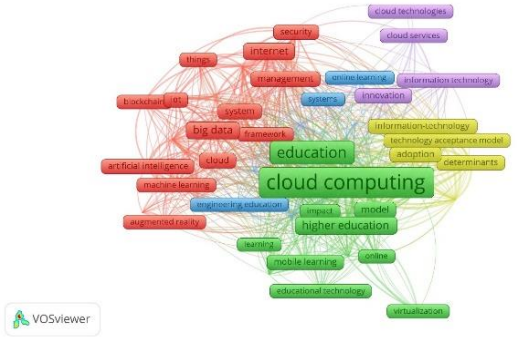


Fig. 3a. Networking for the searchable "Education \ Cloud technologies"

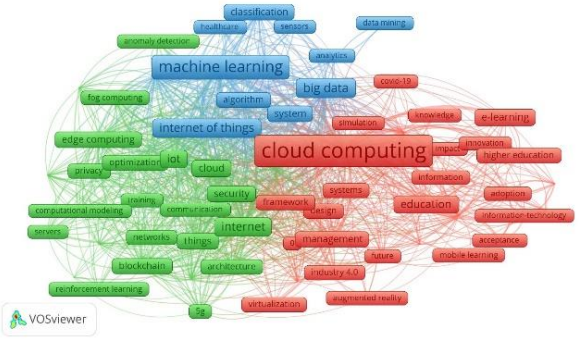


Fig. 3d. Networks general

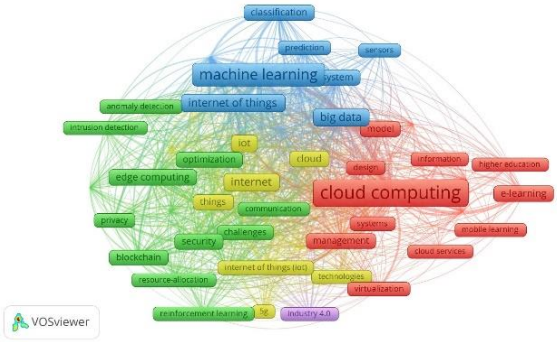


Fig. 3b. Networking for the searchable "Learning \ Cloud technology"

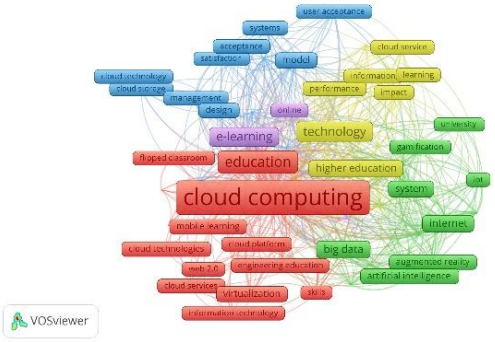


Fig. 3e. Networking for the searchable "Cloud technology\ Teaching"

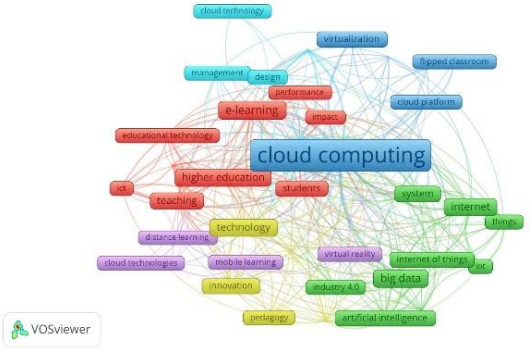


Fig. 3c. Networking for the searchable "Teaching \ Cloud technology"

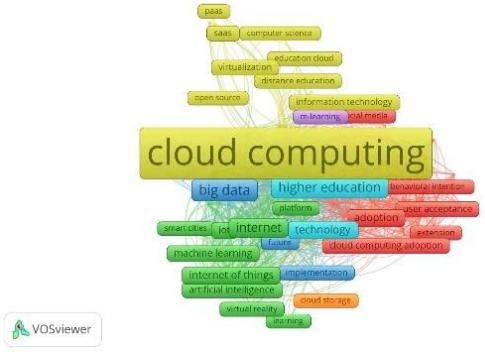


Fig. 3f. Networking for the searchable "Cloud Computing \ Education"

Refinement of the search provides other networks and more detailed accents to characterize the bibliographic landscape of using cloud technologies in education (table 1). The figure (Fig. 2) shows the number of relevant publications by year.

Analysis of the number of published studies in the Web of Science database shows that:

– Over the past ten years, the number of scientific results has increased significantly (almost ten times) – at the beginning of 2014, there were 254 publications on cloud technologies in teaching, and at the beginning of 2023 – 20381. We explain this by the robust development of cloud technologies and their demand in education in the last five years (including during a pandemic);

– The most popular are studies related to learning as one of the directions of cloud technology development – more than 4 thousand publications over the past 20 years. At the same time, the use of cloud technologies by teachers in the educational process is not quantitatively distinguished from other similar searches, which indicates the widespread use of cloud technologies (teachers and students, educational process management, organization of control, independent work, project collaboration, etc.);

– A relatively equal number of studies is presented on education in general as a direction for the development of cloud technologies and on the use of cloud technologies in education – 1662 and 1902 publications, respectively. That indicates the simultaneous interest of educators (in the use of such technologies in educational activities) and IT specialists who offer separate developments/solutions for the field of education;

– Almost the same number of studies is recorded for cloud technologies and teaching: teaching using cloud technologies is described in 1012 publications against 1117 publications related to the development of cloud educational solutions.

– The analysis of the most used keywords and clusters in the built networks (the last column of table 1, Fig. 3) revealed trends in the use of cloud technologies in education:

- research various types of training (mobile, distance, electronic, online) based on cloud technologies;
- research on the impact of Industry 4.0 technologies (Internet of Things (IoT) and 5G technologies, big data, and the Internet) on education and its development;
- research on the use of artificial intelligence in teaching, research on machine learning;
- research of educational process management (management), pedagogical design (design, environment), organization of training by the method of "flipped classroom" and gaming learning technologies (gamification);

- exploring the possibilities of virtualization of education and the use of different types of reality "virtual reality, augmented reality" for training;
- research of problems of visualization of educational material based on cloud technologies and services, which confirm the keywords "performance, model, modeling, simulation".

So, we have identified the following characteristics of the landscape of cloud technology use in education: over the past five years, a significant digital transformation of education due to the growing popularity of cloud technologies in education, teaching, and learning; increasing the number of studies related to various aspects of educational activities under the influence of Industry 4.0 (training, management, design, learning technologies, etc.); a gradual increase in the number of studies on the virtualization of the educational process and the use of artificial intelligence in education; expanding research on the effectiveness of various types of training and teaching methods based on cloud services; the relevance of the trend of visualization of educational material and visual analysis in education.

3. Discussion

The presented results are consistent with the review studies of the use of information and digital technologies in education in recent years. The digital transformation of education due to the growing popularity of cloud technologies in education is recorded by many other publications on various aspects of the implementation of educational activities. Article [12] explores the virtual image of universities. The authors emphasize that today only one in four universities has a balanced digital strategy and prove that digital transformation can provide high-quality and competitive education based on advanced analytics, cloud technologies, and artificial intelligence. Another aspect of the digital transformation of education is presented in [13], which analyzes the development of smart campuses and universities under the influence of key technologies of society. The article describes communication architectures and explores the use of blockchain in university educational programs. The paper [23] is devoted to analyzing the concept of a smart city, in which one of the leading places is occupied by SMART education based on advanced computing technologies (IoT, artificial intelligence, blockchain, big data, and cloud computing). The review of research in [8] addresses the problems of integrating Blockchain and IoT technologies and demonstrates the development of reliable distributed applications in smart education. Another aspect of the digital transformation of education, the use of social networks, is analyzed in the article [6]. The generalization of more than 700 publications on the use of Facebook confirmed this trend. It showed that the use of social media in education is due to external factors and behavioral intentions of subjects of the educational process (user, social and technological aspects). The authors of the publication [39], according to the analysis of more than 1 thousand sources and keywords, identified the educational perspectives of digital learning, including online learning and the development of digital learning environments built using cloud technologies.

The impact of Industry 4.0 on various aspects of the educational process (management, design, learning technologies, etc.) is confirmed by the review works of other scientists. Thus, in [2], it was concluded that Industry 4.0 positively influenced the number and quality of training courses that are distributed online and are considered massive. The authors note the perspective of machine learning, blockchain, and gamification technology to provide students with a richer individual experience. Another aspect of education development under the influence of Industry 4.0 is described in [16]. The authors note society's demand for the formation of autonomous work skills and therefore emphasize the importance of using simulation, horizontal and vertical integration, augmented reality, cybersecurity, and big data and analytics in university professional training programs. The problem of assessing academic progress is raised in [41]. The authors, after analyzing the available scientific results, taking into account the widespread use of digital technologies of Industry 4.0 (big data, cloud computing, and artificial intelligence) in higher education, predict a new approach to the assessment system – the transition to interactive transfer\assessment of knowledge with the organization of space for the exchange of ideas.

The scientific papers also show a gradual increase in the number of studies on using artificial intelligence in learning. The article [5] analyzes the impact of artificial intelligence on higher education. The authors conclude: artificial intelligence is in its infancy for education, but in the future, it will influence the solution of educational problems. In another paper [14], the authors analyze the use of ChatGPT in education, note the ambiguity of perception of this tool by teachers, and at the same time prove that the emergence of AI has already influenced the educational space and, therefore more in-depth studies of its use are needed. The impact of AI, machine learning, and deep learning methods on the organization of the educational process (admission to university, scheduling, and creating courses) is analyzed in [19]. According to the analysis results, 195 original scientific authors raise ethical issues about using artificial intelligence in education.

The educational process's virtualization trend is justified in [26]. The review finds results that show that the research community is most interested in new learning environments, collaboration platforms, and virtual labs. Other work [3] confirms the potential of virtual reality (VR) technologies in developing educational technologies and the organization of educational activities. At the same time, the authors state that the use of VR technologies in architectural and design education still needs to be improved. The article [26] provides an overview of innovative methods, techniques, and learning tools based on cloud computing. The authors prove the promise of such approaches but note the small number of relevant empirical studies (only 17% of 940 publications are based on empirical data). In [24], the Technology Acceptance model was investigated. Based on the analysis of more than 100 publications, the authors stated the need for more internal motivation of the subjects of the educational process to use digital solutions. They confirmed the need for further study of mixed teaching methods.

The expansion of research on the effectiveness of various types of training and teaching methods based on cloud services as a trend is also consistent with the findings of many publications. Thus, in [37], according to the analysis of 429 research articles, it was proved that cloud technologies could become the basis for distance learning and positively affect the quality of education in developing countries. Another paper [4] summarizes scientific findings related to hybrid learning. The authors discuss the revolutionary nature of artificial intelligence to offline and online education through a radical change in communication between the subjects of the educational process and a fundamentally new vision of building an academic environment. The publication [15] provides an overview of the scientific results of applying information technology in vocational education. The review results suggest that using ICTs for vocational education institutions is negligible, especially for monitoring and evaluating learning outcomes. The authors emphasize the need to study robotics, data sciences, artificial intelligence, and cloud computing. The work [31] is devoted to improving the methodology of teaching English. The authors concluded that AI positively impacts learning outcomes by optimizing English language skills, translation, assessment, recognition, attitude, and satisfaction. The studies noted the feasibility of using machine learning, neural networks, genetic algorithms, deep learning, data mining, cloud computing, etc.

The trend of visualization and visual analysis in education is consistent with the conclusions of other studies. Thus, the article [35] is devoted to analyzing trends associated with the visual representation of big data. The authors emphasize that virtual and augmented reality create a new basis for visualization, presentation, and understanding of data. In [17], results were obtained based on visual analytics (visual data, visual networks, and word clouds). The effectiveness of information visualization is proved in the article [18], where the authors demonstrate the effectiveness of using visual materials created based on cloud services. The study [7] analyzes the links between visual and computer thinking. The authors, on the example of contemporary art teaching, show that it can be based on algorithmization and programming principles within the STEAM education framework. The development of visual thinking today can develop using cloud technologies, as noted in [35], and the importance of developing visualization skills for teachers is justified in [36].

Teachers are only sometimes ready to use cloud technologies in their professional activities. Preparing future teachers to use cloud technologies and services is often slowed down by the need for more access to the Internet and the unsatisfactory technical condition and level of computerization of educational institutions. The UNESCO "Global education monitoring report, 2023: technology in education: a tool on whose terms?" [36] is noted that only 10% of high school students use cloud technology for more than an hour a week. The same report notes that the amount of educational content is increasing, but there are no factors that ensure its quality. The report's authors note that society needs more time to accumulate sufficient factual data, which would confirm the nature of the impact of technology on education.

4. Conclusions

The analysis gives grounds for some conclusions.

The development of information technology led to the emergence of cloud computing, which qualitatively influenced the educational industry. Due to quarantine restrictions, cloud technologies have become the driving force for the development of academic learning technologies and have actualized the development of informal and non-formal education. Thanks to the development of digital technologies, mass online courses are more accessible and provide adaptive learning. The increase in online courses will inevitably lead to their new quality. Therefore, we consider the perspective direction of research that will be associated with the development of mass online courses and learning technologies for them:

- immersive technologies;
- the use of virtual, augmented, and mixed reality;
- gaming learning technologies;
- the BYOD approach.

As of today, they have been studied fragmentarily and require additional study and systematization.

Industry 4.0 will accelerate the increasing virtualization of universities, the provision of online learning, and the digitalization of management and organizational decisions at the administrative level. An important aspect will be the continuity of education along with the digital security of each subject of the educational process, so research on the use of blockchain technologies will be relevant. Increased interest in artificial intelligence, machine, and deep learning will determine a new quality in the individualization of the educational process. Inclusive education will develop, and academic analytics and assessment of personal academic results (formative assessment and adaptive computer assessment based on individual results) will become relevant

Education today requires digital skills of the 21st century for everyone, and therefore the current trend will be the advanced training of teachers to use cloud technologies. Educational and professional teacher training programs should include mastery of cloud technologies. It will be essential to develop ideas about cloud technologies in the subject area. Therefore, it will be a perspective to study the use of specialized services in subject education. The spread of artificial intelligence in education will lead to the study of the effectiveness of relevant teaching methods. Therefore, the actual trend is for teachers' advanced preparation for the effective use of AI in pedagogical activities. The challenge for working teachers may be the problem of motivation to use cloud technologies in their professional activities. The search for effective online models of teacher internships will become another promising research trend.

Analysis of any processes today faces the problem of processing big data that can be visualized using cloud technologies. Therefore, studies related to their visualization (design, simulation, modeling of various processes, etc.) will remain promising. The need to develop new academic disciplines (for example, "interaction design", "human-computer interaction", and "user experience") for the presentation of various data and design of the dynamics of their interaction is actualized. The trend of visualization in education and the appropriate training of teachers to develop visual images of knowledge models and use them in teaching will remain popular.

Today, the skills of critical, algorithmic, computational, visual thinking, and modeling, which are formed on an interdisciplinary basis, are in demand. Therefore, the trend of STEM and STEAM education research as the development of interdisciplinary connections between natural sciences, mathematics, engineering, and digital technologies will gain popularity. Research in vocational education will also be promising: effective forms, methods, and means of teaching robotics, data sciences, artificial intelligence, and cloud computing with an occupation accent.

References

[1] Adtani R. et al.: Embracing ICT in academia: adopting and adapting to the new normal pedagogy. Global Knowledge Memory and Communication, 2023 [http://doi.org/10.1108/GKMC-03-2023-0089].

[2] Ahmad I. et al.: MOOC 5.0: A Roadmap to the Future of Learning. Sustainability 14(18), 2022, 11199 [http://doi.org/10.3390/su141811199].

[3] AL-Kebsi S. I., Mostafa A. O. M. S.: Trends and Challenges of Virtual Reality in Architectural Design Education. Journal of Architecture and Planning-king saud university 33(2), 2021, 191-215 [http://doi.org/10.33948/JAP-KSU-33-2-3].

[4] Almusaed A. et al.: Enhancing Student Engagement: Harnessing "AIED"'s Power in Hybrid Education-A Review Analysis. Education Sciences 13(7), 2023, 632 [http://doi.org/10.3390/educsci13070632].

[5] Alotaibi N. S., Alshehri A. H.: Prospers and Obstacles in Using Artificial Intelligence in Saudi Arabia Higher Education Institutions-The Potential of AI-Based Learning Outcomes. Sustainability 15(13), 2023, 10723 [http://doi.org/10.3390/su151310723].

[6] Al-Qaysi N. et al.: Social media adoption in education: A systematic review of disciplines, applications, and influential factors. Technology in Society 73, 2023, 102249 [http://doi.org/10.1016/j.techsoc.2023.102249].

[7] Buhl M.: Computational Thinking Utilizing Visual Arts, or Maybe the Other Way Around. Proceedings of the 18th European Conference on E-learning, 2019, 102-108 [http://doi.org/10.34190/EEL.19.138].

[8] Chen Y. et al.: Applications of Blockchain in Industry 4.0: a Review. Inf Syst Front, 2022 [http://doi.org/10.1007/s10796-022-10248-7].

[9] Dron V.: Formation of research competencies in students during computer modeling of physical phenomena and processes in distance learning. Physical and Mathematical Education 35(3), 2022, 19-25 [https://doi.org/10.31110/2413-1571-2022-035-3-003].

[10] Drushlyak M. et al.: Use of specialized software for the development of visual thinking of students and pupils. Innovative Educational Technologies, Tools, and Methods for E-learning. Scientific Editor Eugenia Smyrnova-Trybulska "E-learning", 12, Katowice–Cieszyn 2020, 147–158 [http://doi.org/10.34916/el.2020.12.13].

[11] Drushlyak M.G. et al.: The Automated Control of Students Achievements by Using Paper Clicker Plickers. International conventions on Information and communication technology, electronics and Microelectronics – MIPRO 2020, 2020, 688-692 [http://doi.org/10.23919/MIPRO48935.2020.9245281].

[12] Fernandez A. et al.: Digital transformation initiatives in higher education institutions: A multivocal literature review. Educ Inf Technol 28, 2023, 12351–12382 [http://doi.org/10.1007/s10639-022-11544-0].

[13] Fernandez-Carames T. M., Fraga-Lamas P.: Towards Next Generation Teaching, Learning, and Context-Aware Applications for Higher Education: A Review on Blockchain, IoT, Fog and Edge Computing Enabled Smart Campuses and Universities. Applied sciences-Basel 9(21), 2019, 4479 [http://doi.org/10.3390/app9214479].

[14] Grassini S.: Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. Education Sciences 13(7), 2023, 692 [http://doi.org/10.3390/educsci13070692].

[15] Hassan R. H., Hassan M. T., Naseer S., Khan Z., Jeon M.: ICT Enabled TVET Education: A Systematic Literature Review. IEEE Access 9, 81624-81650 [http://doi.org/10.1109/ACCESS.2021.3085910].

[16] Hernandez-de-Menendez M., Diaz C. E. A., Morales-Menendez R.: Engineering education for smart 4.0 technology: a review. International Journal of interactive design and manufacturing 14(3), 2020, 789-803 [http://doi.org/10.1007/s12008-020-00672-x].
[17] Herrada R. I., Banos R., Alcayde A.: Student Response Systems: A Multidisciplinary Analysis Using Visual Analytics. Education Sciences 10(12), 2020, 348 [http://doi.org/10.3390/educsci10120348].
[18] Liu Z. J., Levina V., Frolova, Y.: Information Visualization in the Educational Process: Current Trends. International Journal of emerging technologies in learning 15(13), 2020, 49–62 [http://doi.org/10.3991/ijet.v15i13.14671].
[19] Mallik S., Gangopadhyay A.: Proactive and reactive engagement of artificial intelligence methods for education: a review. Frontiers in artificial intelligence 6, 2023, 1151391 [http://doi.org/10.3389/frai.2023.1151391].
[20] Mhlongo S. et al.: Challenges, opportunities, and prospects of adopting and using smart digital technologies in learning environments: An iterative review. Heliyon 9(6), 2023, e16348 [http://doi.org/10.1016/j.heliyon.2023.e16348].
[21] Mukul E., Buyukozkan G.: Digital transformation in education: A systematic review of education 4.0. Technological forecasting and social change 194, 2023, 122664 [http://doi.org/10.1016/j.techfore.2023.122664].
[22] Palamarchuk A. S.: Using cloud service OneDrive in the educational process of the university. Physics and Mathematics Education 2(8), 2016, 87–92.
[23] Rani S.: Amalgamation of Advanced Technologies for Sustainable Development of Smart City Environment: A Review. IEEE Access 9, 2021, 150060-150087 [http://doi.org/10.1109/ACCESS.2021.3125527].
[24] Rosli M.S.: A Systematic Review of the Technology Acceptance Model for the Sustainability of Higher Education during the COVID-19 Pandemic and Identified Research Gaps. Sustainability 14(18), 2022, 11389 [http://doi.org/10.3390/su141811389].
[25] Rudenko Yu. et al.: Online Learning with the Eyes of Teachers and Students in Educational Institutions of Ukraine. TEM Journal 10(2), 2021, 922-931 [http://doi.org/10.18421/TEM102-55].
[26] Scalera M., Gentile E., Plantamura P., Dimauro G.: A Systematic Mapping Study in Cloud for Educational Innovation. Appl. Sci. 10(13), 2020, 4531 [http://doi.org/10.3390/app10134531].
[27] Semenikhina E. et al.: Cloud-based service GeoGebra and its use in the educational process: the BYOD approach. TEM JOURNAL – Technology, Education, Management, Informatics 8(1), 2019, 65–72 [http://doi.org/10.18421/TEM81-08].
[28] Semenikhina O. et al.: The Formation of Skills to Visualize by the Tools of Computer Visualization. TEM Journal 9(4), 2020, 1704–1710. [http://doi.org/10.18421/TEM94-51].
[29] Semenikhina O. V., Drushliak M. G., Khvorostina Yu. V.: Use of GeoGebra cloud service in future math teachers' teaching. Information Technologies and Learning Tools 73(5), 2019, 48–66 [http://doi.org/10.33407/itlt.v73i5.2500].
[30] Shamonia V., Semenikhina O., Drushlyak M.: Use of the Proteus for visual modeling of the work of the information system basic elements. Physical and Mathematical Education 2(20), 2019, 160–165 [http://doi.org/10.31110/2413-1571-2019-020-2-025].
[31] Sharadgah T. A., Sa'di R. A.: A systematic review of research on the use of artificial intelligence in English language teaching and learning (2015-2021): what are the current effects? Journal of information technology education-research 21, 2022, 337–377 [http://doi.org/10.28945/4999].
[32] Shyshkina M., Nosenko Yu.: Cloud technologies of open science in the process of continuous training of ICT in education. Physical and Mathematical Education 37(5), 2022, 69–74 [http://doi.org/10.31110/2413-1571-2022-037-5-010].
[33] Smestad B. et al.: Examining dimensions of teachers' digital competence: A systematic review pre- and during COVID-19. Heliyon 9(6), 2023, e16677 [http://doi.org/10.1016/j.heliyon.2023.e16677].
[34] Su J. H., Yang W. P.: Digital competence in early childhood education: A systematic review. Education and information technologies, 2023 [http://doi.org/10.1007/s10639-023-11972-6].
[35] Tamayo J. L. R., Hernandez M. B., Gomez H. G.: Digital Data Visualization with Interactive and Virtual Reality Tools. Review of Current State of the Art and Proposal of a Model. Journal ICONO 14 16(2), 2018 [http://doi.org/10.7195/ri14.v16i2.1174].
[36] Technology in education: a tool on whose terms? Global education monitoring report, UNESCO, 2023 [https://unesdoc.unesco.org/ark:/48223/pf0000385723].
[37] Thavi R. et al.: Role of cloud computing technology in the education sector. Journal of engineering design and Technology, 2021 [http://doi.org/10.1108/JEDT-08-2021-0417].
[38] Turinov A., Galdina A.: Application of computer modeling to solving quantum-mechanical problems. Physical and Mathematical Education 3(13), 2017, 170–177.
[39] Vaicondam Y. et al.: Research Landscape of Digital Learning Over the Past 20 Years: A Bibliometric and Visualisation Analysis. International Journal of Online and biomedical engineering 18(8), 2022, 4–22 [http://doi.org/10.3991/ijoe.v18i08.31963].
[40] Vesic D., Lakovic D., Vesic S. L.: Use of information technologies in higher education from the aspect of management. International journal of cognitive research in science engineering and education 11(1), 2023, 143–151 [http://doi.org/10.23947/2334-8496-2023-11-1-143-151].
[41] Xu X. et al.: Review on A big data-based innovative knowledge teaching evaluation system in universities. Journal of innovation & knowledge 7(3), 2022, 100197 [http://doi.org/10.1016/j.jik.2022.100197].
[42] Yurchenko A. et al.: Using online IT-industry courses in the computer sciences specialists' training. International Journal of Computer Science and Network Security 21(11), 2021, 97–104. [http://doi.org/10.22937/IJCSNS.2021.21.11.13].
[43] Yurchenko A.: Digital physical laboratories as an important means of training of future teachers of physics. Physical and Mathematical Education 1(4), 2015, 55–63.

Ph.D. Artem Yurchenko
e-mail: a.yurchenko@fizmatsspu.sumy.ua

Ph.D. of Pedagogical Sciences, associate professor, Department of Computer Science, Sumy State Pedagogical University named after A.S. Makarenko, Sumy, Ukraine.
Scientific interests: IT in education, computer visualization of knowledge and computer simulation, development of digital competence of future teachers, specialized computer software in the field of natural sciences, and its use.

http://orcid.org/0000-0002-6770-186X

Ph.D. Anzhela Rozumenko
e-mail: angelarozumenko@ukr.net

Ph.D. of Pedagogical Sciences, Associate Professor, Department of Higher Mathematics, Sumy National Agrarian University, Sumy, Ukraine.
Scientific interests: higher mathematics, teaching methods, IT in education, history of mathematics, mathematical statistics.

http://orcid.org/0000-0002-4759-3320

Ph.D. Anatolii Rozumenko
e-mail: a.rozumenko@snau.edu.ua

Ph.D. of Physical and Mathematical Sciences, Associate Professor, Department of Higher Mathematics, Sumy National Agrarian University, Sumy, Ukraine.
Scientific interests: higher mathematics, applied mathematics, probability theory, teaching mathematics, IT in education.

http://orcid.org/0000-0002-3069-9313

M.Sc. Roman Momot
e-mail: r.momot@fizmatsspu.sumy.ua

Postgraduate, Department of Computer Science, Sumy State Pedagogical University named after A.S. Makarenko, Sumy, Ukraine.
Scientific interests: IT in education, computer visualization of knowledge, cloud education technology, cloud services, and its use in education.

http://orcid.org/0000-0003-0861-1925

D.Sc. Olena Semenikhina
e-mail: e.semenikhina@fizmatsspu.sumy.ua

Doctor of Pedagogical Sciences, Professor, Department of Computer Science, Sumy State Pedagogical University named after A.S. Makarenko, Sumy, Ukraine.
Scientific interests: IT in education, computer visualization of knowledge, development of digital competence of future teachers, specialized computer software in the field of natural sciences and its use, and computational mathematics.

http://orcid.org/0000-0002-3896-8151

http://doi.org/10.35784/iapgos.4569 received: 5.09.2023 | revised: 20.10.2023 | accepted: 02.11.2023 | available online: 20.12.2023

HYBRID BINARY WHALE OPTIMIZATION ALGORITHM BASED ON TAPER SHAPED TRANSFER FUNCTION FOR SOFTWARE DEFECT PREDICTION

Zakaria A. Hamed Alnaish[1,2], Safwan O. Hasoon[2]
[1]University of Mosul, College of Science, Mosul, Iraq, [2]University of Mosul, College of Computer Science and Mathematics, Mosul, Iraq

Abstract: *Reliability is one of the key factors used to gauge software quality. Software defect prediction (SDP) is one of the most important factors which affects measuring software's reliability. Additionally, the high dimensionality of the features has a direct effect on the accuracy of SDP models. The objective of this paper is to propose a hybrid binary whale optimization algorithm (BWOA) based on taper-shape transfer functions for solving feature selection problems and dimension reduction with a KNN classifier as a new software defect prediction method. In this paper, the values of a real vector that represents the individual encoding have been converted to binary vector by using the four types of Taper-shaped transfer functions to enhance the performance of BWOA to reduce the dimension of the search space. The performance of the suggested method (T-BWOA-KNN) was evaluated using eleven standard software defect prediction datasets from the PROMISE and NASA repositories depending on the K-Nearest Neighbor (KNN) classifier. Seven evaluation metrics have been used to assess the effectiveness of the suggested method. The experimental results have shown that the performance of T-BWOA-KNN produced promising results compared to other methods including ten methods from the literature, four types of T-BWOA with the KNN classifier. In addition, the obtained results are compared and analyzed with other methods from the literature in terms of the average number of selected features (SF) and accuracy rate (ACC) using the Kendall W test. In this paper, a new hybrid software defect prediction method called T-BWOA-KNN has been proposed which is concerned with the feature selection problem. The experimental results have proved that T-BWOA-KNN produced promising performance compared with other methods for most datasets.*

Keywords: feature selection, binary whale optimization algorithm, taper-shaped transfer function, software defect prediction

HYBRYDOWY, BINARNY ALGORYTM WOA OPARTY NA TRANSMITANCJI STOŻKOWEJ DO PROGNOZOWANIA DEFEKTÓW OPROGRAMOWANIA

Streszczenie: *Niezawodność jest jednym z kluczowych czynników stosowanych do oceny jakości oprogramowania. Przewidywanie defektów oprogramowania SDP (ang. Software Defect Prediction) jest jednym z najważniejszych czynników wpływających na pomiar niezawodności oprogramowania. Dodatkowo, wysoka wymiarowość cech ma bezpośredni wpływ na dokładność modeli SDP. Celem artykułu jest zaproponowanie hybrydowego algorytmu optymalizacji BWOA (ang. Binary Whale Optimization Algorithm) w oparciu o transmitancję stożkową do rozwiązywania problemów selekcji cech i redukcji wymiarów za pomocą klasyfikatora KNN jako nowej metody przewidywania defektów oprogramowania. W artykule, wartości wektora rzeczywistego, reprezentującego indywidualne kodowanie zostały przekonwertowane na wektor binarny przy użyciu czterech typów funkcji transferu w kształcie stożka w celu zwiększenia wydajności BWOA i zmniejszenia wymiaru przestrzeni poszukiwań. Wydajność sugerowanej metody (T-BWOA-KNN) oceniano przy użyciu jedenastu standardowych zestawów danych do przewidywania defektów oprogramowania z repozytoriów PROMISE i NASA w zależności od klasyfikatora KNN. Do oceny skuteczności sugerowanej metody wykorzystano siedem wskaźników ewaluacyjnych. Wyniki eksperymentów wykazały, że działanie rozwiązania T-BWOA-KNN pozwoliło uzyskać obiecujące wyniki w porównaniu z innymi metodami, w tym dziesięcioma metodami na podstawie literatury, czterema typami T-BWOA z klasyfikatorem KNN. Dodatkowo, otrzymane wyniki zostały porównane i przeanalizowane innymi metodami z literatury pod kątem średniej liczby wybranych cech (SF) i współczynnika dokładności (ACC), z wykorzystaniem testu W. Kendalla. W pracy, zaproponowano nową hybrydową metodę przewidywania defektów oprogramowania, nazwaną T-BWOA-KNN, która dotyczy problemu wyboru cech. Wyniki eksperymentów wykazały, że w przypadku większości zbiorów danych T-BWOA-KNN uzyskała obiecującą wydajność w porównaniu z innymi metodami.*

Słowa kluczowe: wybór cech, algorytm optymalizacji binarnej, transmitancja stożkowa, przewidywanie defektów oprogramowania

Introduction

Software defect prediction (SDF) is considered a crucial software quality assurance technique, that could extract defects of any software and help the developers or maintainers efficiently detect the potentially defective modules [30]. To minimize the undesirable effects of any software, software defect prediction should be done before delivering the software to the customers [26]. Generally, building efficient software defect prediction models depends on soft computing (SC), machine learning (ML) [21], software features (metrics) that are generated in the software development process, or code complexity [28]. Besides that, some of these features are unrelated and jobless features.

Consequently, machine learning (ML) techniques consider a good solution for building software defect prediction modules in software projects. Additionally, numerous types of research have confirmed the effectiveness of machine learning (ML) techniques in enhancing the efficiency of software defect prediction methods. The common techniques which are utilized in SDP are SVM, DT, NB, LR, ANN, DF, and CNN [5, 23, 29].

Concurrently with software development and because all life applications are managed by computer systems, and the volume of the produced data will become huge. Thus, the basic ML techniques became unpractical in some fields and need improvements, especially in the SDP field. Dimension reduction considers one of the common methods for enhancing the machine-learning performance of the input data [11]. One of the main techniques of dimension reduction is feature selection (FS) [7, 10].

The feature selection process (FS) involves identifying and choosing the best relevant features for any problem domain to achieve the highest accuracy [7]. The feature selection algorithms are categorized into types [1, 9]. The first type called the filter method, in which the feature selection process does not involve the classifier. The second type called the wrapper method, in which the feature selection relies on the used classifier, which serves as an effective evaluation criterion for choosing the best features. The hybrid method combines wrapper and filter techniques to select the subset of features that rely on the classifier's design [1, 9, 10].

Recently, there are many variants of metaheuristic algorithms that have been implemented for solving feature selection problems in SDP. For instance, in [3] the authors proposed to use fourteen filters as subset feature selection (FSS) methods and four filter feature ranking (FFR). The proposed methods evaluated by utilizing four classifiers based on five types of software defect datasets. The paper compares the performance of feature selection methods, including Relief, Chi-Square, Information Gain, and Correlation-based Feature Selection (CFS). Their results have shown that the efficiency of each method varies according to the dataset type and prediction model.

In [16], the authors proposed a classification framework based on Multi-Layer Perceptron (MLP) and many filter feature selection techniques to predict software defects. The suggested

methods are implemented with and without oversampling techniques for manipulating data misbalancing. The authors have used Twelve datasets with four evaluation metrics. The produced results have demonstrated that the framework with class balancing produces good performance with all datasets.

In [26], the authors proposed a novel approach to improve the performance of a layered-recurrent neural network (L-RNN) for software fault prediction. By using feature selection techniques, the authors aim to eliminate irrelevant features and improve the accuracy of fault prediction. The authors employ three different wrapper feature selection algorithms (Binary Genetic Algorithm, Binary Particle Swarm Optimization, and Binary Ant Colony Optimization) iteratively to select the most important software metrics. The results of the experiments, which are conducted on nineteen common datasets from the PROMISE repository, have shown that the proposed approach achieves an excellent classification rate and outperforms existing results found in the literature. Therefore, the authors claimed that feature selection plays a vital role in enhancing the performance of the layered recurrent neural network for software fault prediction.

In [6], to address the high dimensionality and filter rank selection problem in software defect prediction, the authors suggested a unique rank aggregation-based multifilter FS method. The suggested approaches combine rank lists produced by various filter methods into a single aggregated rank list by employing rank aggregation algorithms. On nine defect datasets from the NASA repository, the efficiency of the suggested technique was assessed using Decision Tree (DT) and Naive Bayes (NB) models. According to the experimental findings, the proposed methods had a greater impact on the prediction performances of NB and DT models than other FS methods.

In [25], a new version of the Binary Moth Flame Optimization called (EBMFO) algorithm and Adaptive synthetic sampling (ADASYN) has been proposed to predict software defects. The algorithm addresses the issue of imbalanced data distribution and enhances the input dataset for accurate predictions. The proposed EBMFO algorithm is employed as a wrapper feature selection algorithm, which selects the most relevant features from the input dataset to enhance the overall performance of classifiers. Also, the authors demonstrated that transfer functions are important in the Enhanced Binary Moth Flame Optimization algorithm as they are used to convert the continuous algorithm to a binary version where eight different transfer functions from two groups, to examine the probabilities of updating the process of choosing features from a binary vector, S-shaped and V-shaped models are adopted. The given results show that the suggested EBMFO improves the classifier's efficiency overall and outperforms the findings in the literature.

In [11], the authors suggested an enhanced version of the Whale Optimization Algorithm (WOA) by incorporating natural selection operators to improve software fault prediction. The authors introduce two natural selection operators: crossover and mutation. The crossover operator facilitates the exchange of genetic information between two whale individuals, while the mutation operator introduces small changes to the solution space. The proposed Boosted Whale Optimization Algorithm (BWOA) is evaluated using multiple real-world software datasets. The experimental results demonstrate that BWOA outperforms traditional WOA and other state-of-the-art fault prediction techniques in terms of accuracy, precision, recall, and F-measure. Thus, the BWOA algorithm considers an efficient method for solving feature selection problems in SDP.

In light of the studies mentioned above, the motivation for this study is raised. Although there were many methods have been proposed recently for solving feature selection problems in SDP, there is still a need to develop and enhance a robust feature selection method because not all the existing feature selection techniques produce the best accuracy in solving SDP with all datasets. Since the transfer functions dynamically adjust the search behaviour based on the fitness values of the solutions and allow for more efficient exploration of the search space [12]. Thus, using an appropriate transfer function to convert the continuous search space to binary search space considers a big challenge. So, introducing a new transfer function considers the objective of this study.

Thus, this paper proposes a new robust feature selection method based on BWOA and taper-shaped transfer function. The following points summarized the contributions of this paper:

1) Propose a new hybrid feature selection method using a binary whale optimization algorithm and Taper Shaped Transfer Function to get the most significant features for solving software defect prediction problems.
2) The proposed method (T-BWOA-KNN) has been evaluated based on eleven SDP datasets with (KNN) classifier in terms of accuracy rate, average of selected features number, sensitivity, specificity, accuracy, G-mean, and error rate (ER)
3) Kandell W test has been implemented to show the significant difference for the performance of the (T-BWOA-KNN) and rank it with the techniques from the literature.

The rest of this paper is arranged as follows: section two includes an explanation of the binary whale optimization algorithm (BWOA). Section three explains the suggested method. The experiment setup has been explained in section four. Section five includes the experiment results and discussions. Section Six includes a comparison performance of the proposed T-BWOA-KNN with other methods from the literature. Finally, section seven includes the conclusion.

1. Binary Whale Optimization Algorithm (BWOA)

First, Whale Optimization Algorithm (WOA) [20] is an optimization algorithm that is inspired by the foraging behaviour of whales. Whales exhibit a feeding technique known as the bubble-net method during their foraging activities. Nevertheless, within the Whales Optimization Algorithm, the instant most promising solution is designated as either the desired prey or positioned in close proximity to the optimal solution. The remaining whales then strive to adjust their positions towards this superior choice. For solving any optimization problem, the implementation of the WOA algorithm could be represented as n_c represents the number of Whales, and x_i^t represents the value for the position of the (i*th*) whale at iteration (*t*). For instance, assume at iteration *t*, for *n* whales and *d* dimensions, the whales are denoted as shown in matrix (*X*):

$$whales(X) = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_d^1 \\ x_1^2 & x_2^2 & \dots & x_d^2 \\ \vdots & \vdots & \dots & \vdots \\ x_1^n & x_2^n & \dots & x_d^n \end{bmatrix}$$

Each row in the matrix (X) denotes one probable solution.

The mathematical simulation process of WOA swarming behavior could be described as shown in the following equations:

$$D = |C.X^*(t) - X\,(t)\,| \quad (1)$$

$$X(t+1) = X^*(t+1) - A.D \quad (2)$$

where *t* represents the iteration number, *X* represents the position vector, and X^* represents the position vector of the best-founded solution. A and C represent the coefficient vectors as shown in equations 3 and 4, respectively:

$$A = 2ar - a \quad (3)$$

$$C = 2r \quad (4)$$

The variable "*r*" is assigned a random value within the range of [0, 1], while the variable "a" undergoes a linear decrease from two to zero throughout the repeated cycles. Similar to other optimization algorithms, this algorithm consists of two main phases: exploration and exploitation. The exploitation phase involves two processes:

Shrinking encircling mechanism: This mechanism is achieved by reducing the value of "*a*" based on equation 4. It is important to note that "*a*" is a random value within the range of [-*a*, *a*].

Spiral updating position: this process includes calculating the distance between the whale and the prey. Then, the spiral

equation is utilized to imitate the movement resembling a helix, as shown in equation 5.

$$X(t+1) = D^1 e^{bl} \cos(2\pi l) + X^*(t) \quad (5)$$

where l represents a random number within [−1, 1] and b is a constant. It is assumed that there is a 50% chance of selecting either the shrinking encircling mechanism or the spiral model. It can be represented mathematically as shown in equation 6:

$$X(t+1) = \begin{cases} X^*(t) - AD & if\ p < 0.5 \\ D^1 e^{bl} \cos(2\pi l) + X^*(t) & if\ p \geq 0.5 \end{cases} \quad (6)$$

where p is a uniformly distributed random number. In the exploration phase, to encourage the agent to move away from its current position, random values within the range of 1 < A < -1 are utilized, as shown in equations 7 and 8:

$$\mathrm{D} = |\mathrm{CX_{rand}} - \mathrm{X}| \quad (7)$$

$$\mathrm{X(t+1)} = \mathrm{X_{rand}} - \mathrm{AD} \quad (8)$$

In the binary WOA (BWOA) [19], whales move inside a binary search space instead of a continuous search space in order to modify their positions. So, to solve the feature selection problem, the solutions must be represented as 0 and 1 only. Actually, there are two versions of BWOA which are S-BWOA and V-BWO based on the used transfer functions, either the S-shaped function or the V-shaped function [14].

2. The proposed method (T-BWOA) for software defect prediction

In the binary Whale Optimization Algorithm (BWOA), the solutions are represented by binary space [7]. In feature selection problem, the continuous spaces should be transforming to the equivalent binary space (i.e. 0 or 1), where the value 0 means that the feature is irrelevant and the value 1 means that the feature is relevant.

Also, in the BWOA, the transformation process needed using a appropriate transfer function such as sigmoid function (S-shaped) and hyperbolic function (V-shaped). The transfer function plays an important role in the performance of the binary optimization algorithm (BOA) [12]. Hence, the motivation of this study is to propose a new hybrid BWOA called taper shaped transfer function-binary whale optimization algorithm (T-BWOA) for solving feature selection problem in software defect prediction. In this study, four taper-shaped transfer functions (T1-T4) [12] are used to convert the continues values (x) to real number within [0,1] based on the equations (9–12):

$$T1(x) = \frac{\sqrt{|x|}}{\sqrt{|A|}} \quad (9)$$

$$T2(x) = \frac{|x|}{|A|} \quad (10)$$

$$T3(x) = \frac{\sqrt[3]{|x|}}{\sqrt[3]{A}} \quad (11)$$

$$T4(x) = \frac{\sqrt[4]{|x|}}{\sqrt[4]{|A|}} \quad (12)$$

Figure (1) shows the curves of Taper-shaped transfer functions.

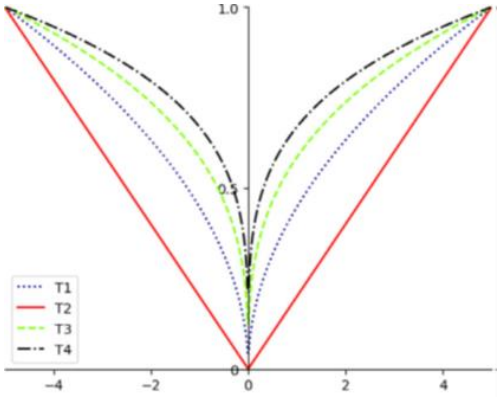


Fig. 1. The curves of Taper-shaped transfer functions [12]

Then, convert the obtained values by $Tk(x)(k = 1, 2, 3, 4)$, to the binary space based on equation (13)

$$x_b = \begin{cases} 1, & if\ T_k(x) \leq T_s \\ 0, & otherwise \end{cases} \quad (13)$$

Then, the fitness value is calculated using a fitness function which is represented by equation (14)

$$f\left(x_{i,t}\right) = \propto C_r\left(x_{i,t}\right) + \beta \frac{|X|}{|K|} \quad (14)$$

where C_r represents the error rate of the classification, X represents the number of selected features by whale $(x_{i,t})$, K represents all the features, $\propto$ =0.99 and β =1-$\propto$. Algorithm 1 shows the steps of the proposed T-BWOA.

Algorithm 1: **(T-BWOA-KNN)**
1: **start**
2: **The inputs**: a number of whales (N) and maximum number of iterations (t).
3: **The output**: the best whale's positions
4: set the Initial values of a and N
5: compute the fitness value for the whales according to equation (4) and find the best search agent (X^*)
6: **while** stop condition**(** maximum iteration (t)) has not satisfied do
7: **for** i = 1: N
8: Calculate and Update the following parameters:
9: a = 2 - t * (2 / t)
10: A = 2 * a * rand() – a
11: C= 2 * rand()
12: P= rand(),
13: L= random number in [-1,1]
14: ***if*** *p >= 0.5 then*
15: $X(t+1) = D' * e^{b*l} * \cos(2\pi l) + x^*(t)$
16: $D' = x^*(t) - x(t)$
17: **Else** (p<0.5) do
18: ***if*** $(|A| < 1)$***then***
19: $x(t+1) = x^*(t) - A * D$
20: $D = |C * x^*(t) - x(t)|$
21: **Else** $(|A| \geq 1)$
22: $x(t+1) = x_{rand} - A * D_x$
23: $D_x = |C * x_{rand} - x|$
24: **End if**
25: **End if**
26: **Convert** $x(t+1)$ **to binary** space *using four types of taper shaped transfer functions* based on equation (13) for each t value
27: **End for**
28: calculate the fitness value using **KNN** classifier for each whale based on equation (14)
29 Update the value of x^*, if there is a better solution
30: t=t+1
31: **end while**

Also, figure 2. demonstrates the framework of the suggested T-BWOA-KNN.

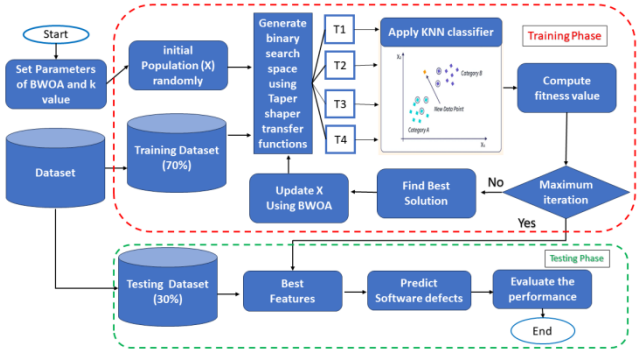


Fig. 2. The framework of the proposed T-BWOA-KNN

3. Experiments setup

In this section, the datasets description, the pre-processing, performance evaluation metrics, parameters configuration, and the used statistical test are explained in details in the following subsections. The results of each experiment are obtained by running each experiment ten times separately then calculate the average of all experiments.

3.1. Datasets description

To measure the performance of the proposed method (T-BWOA-KNN), eleven software datasets were used which are available free in NASA and PROMISE repository [2, 17, 22]. Each dataset includes different number of records. Each record represents one project. Also, each project has many features (attribute).

Each dataset has divided into two groups which are: training part consists of 70% of the total size of the dataset, while the remaining instances form the testing part. Each experiment is independently conducted 10 times. Table 1 shows a summary of the used dataset.

Table 1. Datasets description

Dataset	Attribute (features)	No. of records	Defective records	Non-defective records
CM1	37	327	42	285
JM1	21	9593	1759	7834
KC1	21	2096	325	1771
KC3	39	194	36	158
MW1	37	250	25	225
PC1	37	679	55	624
PC2	36	1585	16	1569
PC3	37	1125	140	985
PC4	37	1270	176	1094
PC5	39	17186	516	16670
Tomcat	20	858	77	781

3.2. Pre-processing

Min-Max scaler normalization has been applied in this paper because it is a normalization technique commonly used in machine learning and data preprocessing. It scales the features of a dataset to a specific range, typically between 0 and 1 [27], as shown in equation (15).

$$\text{Normalized value } (x') = (x - \min)/(\max - \min) \quad (15)$$

3.3. Performance evaluation

Predicting the defective classes in a target version based on a confusion matrix, as shown in table 2.

Table 2. Confusion matrix

Predicted output	Actual output	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Where: TP is true positive, TN is true negative, FP is false positive, FN is false negative. In this research, based on the values of these four indicators, seven evaluation metrics were calculated to measure the performance of the suggested method which are: classification accuracy (ACC), area under carve (AUC), sensitivity (SN), specificity (SP), number of selected features (SF), error rate (ER), and G-mean.

The classification accuracy (ACC) represents the ratio of the instances that have been classified correctly [30], it is calculated using equation (16).

$$ACC = (TP + TN) / (TP + TN + FP + FN) \quad (16)$$

Also, area under carve (AUC) has been used to assess the distinguishing ability of the proposed model. its value falls within [0,1], the higher the better. In addition, AUC is appropriate for evaluating class-imbalanced datasets [8] . it is calculated based on equation (17)

$$AUC=(1+TPR-FPR)/2 \quad (17)$$

where TPR is the proportion of positive label instances that were predicted correctly [21], as shown in equation (18).

$$TPR=TP/ (TP+ FN) \quad (18)$$

And, false positive rate (FPR) represents the proportion of negative instances that are incorrectly predicted as positive by the model, which is calculated as shown in equation (19).

$$FPR=FP/(FP+TN) \quad (19)$$

The false positive rate is typically expressed as a percentage or a decimal value between 0 and 1. A lower FPR indicates a more accurate model, as it means fewer negative instances are being misclassified as positive.

In addition, G-mean is used to show the efficiency of both sensitivity (SN) and specificity (SP) together [10], which is calculated using equation (20).

$$G\text{-mean}= \sqrt{(SN*SP)} \quad (20)$$

where SN is the probability of correct classification for the positive instances. SP represents the probability of correct classification for the negative instances, which are calculated based on equations (21) and (22), respectively.

$$SN=TP /(TP +FN) \quad (21)$$

$$SP= TN/(TN +FP) \quad (22)$$

In addition, the error rate (ER) has been calculated based on equation (23), which shows the misclassification rate for the classes.

$$ER=1- \text{Accuracy} \quad (23)$$

3.4. Parameters configuration

The values of the parameters for each experiment are set as shown in table 3.

Table 3. The values of parameters for each experiment

Parameter's name	Parameters values
Number of whales	120
Number of iterations	100
Number of dimensions	Number of total features for each dataset
Initial values	[0,1]
b	1
K	5

3.5. Statistical test

In this research, Kendall W test has been used to show the significant performance of the suggested techniques and rank it with other SDP methods from the literature. The Kendall W test is a statistical test used to measure the degree of agreement among multiple observers or raters. It assesses the extent to which the rankings or ratings assigned by different observers to a set of items or subjects are consistent. In the Kendall W test, each case represents a judge or rater, while each variable represents the thing or person being assessed, below are the steps for calculating the Kendall W test score [19]:

1. Assume the object (i) is considered as the SDP method, (ranked objects) is given the rank rij by the raters j (datasets), where there are in total (n) objects and (m) raters. The total rank (R) given to object (i) is calculated using equation (24)
$$R_i = \sum_{j=}^{m} r_{i,j} \quad (24)$$
2. The average value of (R) is calculated using equation (25)
$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i \quad (25)$$
3. The sum of squared deviations, S, is calculated using equation (26)
$$S = \sum_{i=1}^{n}(R_i - \bar{R})^2 \quad (26)$$
4. The Kendall's W coefficient is calculated using equation (27)
$$W = \frac{12*S}{m^2(n^3-n)} \quad (27)$$

The score's range of the Kendall W test will be within the interval [0,1] and the decision will depend on the following roles[28] :

- 0.00 <= w < 0.20 – Slight agreement
- 0.20 <= w < 0.40 – Fair agreement
- 0.40 <= w < 0.60 – Moderate agreement
- 0.60 <= w < 0.80 – Substantial agreement
- w >= 0.80 – Almost perfect agreement

4. The experiments results and discussions

The final implementation of the proposed T-BWOA with KNN classifier is done in Python 3.9.7 using Spyder which is served as the development environment. The experimental results for the eleven datasets are provided in tables 4–10. These tables show the obtained results for the suggested method (T-BWOA-KNN) for solving SDP problem in terms of accuracy rate (ACC), the average number of selected features (SF), sensitivity (SN), specificity (SP), area under carve (AUC), G-mean, and error rate (ER). Every table shows the values of one evaluation metric and each row in the table displays the obtained result for one dataset, with the best result shaded for each method whereas the first column represents name of the dataset and the rest represents the proposed methods.

In terms of accuracy, As shown in table 4, T2-BWOA has produced the highest accuracy for JM1, KC1, PC2, PC3 datasets. Meanwhile, T1-BWOA has produced the highest accuracy for CM1, JM1, PC1 datasets. T3-BWOA has produced the highest accuracy for MW1, PC4, TOMCAT datasets. T4-BWOA has produced the highest accuracy for JM1, KC3, PC5 datasets. In summary, T2-BWOA has the best performance in terms of accuracy for four from eleven datasets comparing with others whose has the best performance for three datasets only which are shaded by gray color; although T1-BWOA and T3-BWOA has produced the best performance in terms of accuracy mean (0.891%).

In terms of sensitivity, as shown in table 5, T1-BWO, and T3-BWOA have produced the highest sensitivity for four different datasets which are shaded by gray colour. Meanwhile, T4-BWOA has produced the highest sensitivity for only two datasets. At the same time, T2-BWOA has produced the highest performance in terms of sensitivity mean. Although T2-BWOA has the best performance in terms of sensitivity for three datasets from eleven, the T2-BWOA has produced the best performance in terms of sensitivity mean (0.296%).

In terms of Specificity, as shown in table 6, T2-BWOA has produced the highest specificity for five datasets from eleven. Meanwhile, T1-BWOA has produced the highest specificity for PC1 dataset only. T3-BWOA has produced the highest specificity for four datasets. T4-BWOA has produced the highest specificity for three datasets which are shaded by gray color. In summary, T2-BWOA has the best performance in terms of specificity for five datasets from eleven comparing with others. Also, the T2-BWOA has produced the best performance in terms of specificity mean (0.964%).

In terms of G-mean, as shown in table 7, although T1-BWOA, T3-BWOA, and T4-BWOA have produced the highest G-mean for three different datasets which are shaded by gray colour. The T4-BWOA consider the best in terms of (G-mean) mean (0.456%).

In terms of AUC, as shown in table 8, T4-BWOA has produced the highest AUC for four datasets from eleven. Meanwhile, T1-BWOA has produced the highest AUC for three datasets. T2-BWOA has produced the highest AUC for PC2 dataset only. T3-BWOA has produced the highest accuracy for three datasets. In summary, T4-BWOA has the best performance in terms of AUC for four datasets from eleven datasets compared with others which are shaded by gray color. Also, the T4-BWOA has produced the best performance in terms of AUC mean (0.6257%).

In terms of error rate, as shown in table 9, T2-BWOA has produced the minimum error rate for four datasets from eleven. Meanwhile, T1-BWOA, T3-BWOA, and T4-BWOA has produced the minimum error rate for three datasets which are shaded in gray colour. Although, T3-BWOA has produced the best performance in terms of mean error rate (0.1080).

Table 4. The obtained results of T-BWOA in terms of accuracy (%)

Dataset	T1-BWOA	T2-BWOA	T3-BWOA	T4-BWOA
Cm1	0.851	0.849	0.84	0.841
JM1	0.958	0.958	0.956	0.958
KC1	0.846	0.850	0.842	0.842
KC3	0.790	0.781	0.786	0.798
MW1	0.896	0.86	0.900	0.888
PC1	0.913	0.902	0.901	0.884
PC2	0.988	0.990	0.989	0.988
PC3	0.853	0.868	0.852	0.861
PC4	0.862	0.843	0.869	0.867
PC5	0.973	0.974	0.976	0.977
tomcat	0.879	0.892	0.900	0.89
Mean	0.891	0.887	0.891	0.890
Total shaded	3	4	3	3

Table 5. The obtained results of T-BWOA in terms of sensitivity (%)

Dataset	T1-BWOA	T2-BWOA	T3-BWOA	T4-BWOA
Cm1	0.084	0.038	0.031	0.046
JM1	0.866	0.862	0.861	0.858
KC1	0.368	0.366	0.400	0.347
KC3	0.127	0.136	0.127	0.164
MW1	0.238	0.175	0.263	0.2
PC1	0.194	0.131	0.194	0.238
PC2	0.02	0.100	0.02	0.100
PC3	0.260	0.190	0.183	0.214
PC4	0.283	0.038	0.315	0.262
PC5	0.429	0.862	0.479	0.512
tomcat	0.178	0.366	0.161	0.226
Mean	0.277	0.296	0.275	0.287
Total shaded	4	3	4	2

Table 6. The obtained results of T-BWOA in terms of specificity (%)

Dataset	T1-BWOA	T2-BWOA	T3-BWOA	T4-BWOA
Cm1	0.968	0.973	0.964	0.962
JM1	0.978	0.980	0.978	0.980
KC1	0.934	0.940	0.923	0.933
KC3	0.945	0.932	0.940	0.947
MW1	0.975	0.942	0.976	0.970
PC1	0.974	0.968	0.962	0.939
PC2	0.999	1	1	0.998
PC3	0.938	0.964	0.947	0.953
PC4	0.955	0.946	0.958	0.964
PC5	0.990	0.991	0.992	0.991
tomcat	0.948	0.970	0.972	0.955
Mean	0.964	0.964	0.964	0.962
Total shaded	1	5	4	3

Table 7. The obtained results of T-BWOA in terms of G-mean(%)

Dataset	T1-BWOA	T2-BWOA	T3-BWOA	T4-BWOA
Cm1	0.210	0.101	0.091	0.128
JM1	0.920	0.919	0.918	0.917
KC1	0.579	0.580	0.604	0.561
KC3	0.300	0.318	0.259	0.316
MW1	0.417	0.396	0.465	0.434
PC1	0.335	0.301	0.360	0.398
PC2	0.045	0.197	0.045	0.171
PC3	0.484	0.420	0.408	0.444
PC4	0.494	0.419	0.547	0.493
PC5	0.647	0.629	0.689	0.710
tomcat	0.354	0.255	0.354	0.451
Mean	0.435	0.412	0.430	0.456
Total shaded	3	2	3	3

Table 8. The obtained results of T-BWOA in terms of AUC (%)

Dataset	T1-BWOA	T2-BWOA	T3-BWOA	T4-BWOA
Cm1	0.526	0.506	0.497	0.504
JM1	0.922	0.921	0.920	0.919
KC1	0.651	0.653	0.662	0.640
KC3	0.536	0.534	0.534	0.555
MW1	0.606	0.558	0.619	0.585
PC1	0.584	0.549	0.578	0.588
PC2	0.509	0.550	0.51	0.549
PC3	0.599	0.577	0.565	0.583
PC4	0.619	0.575	0.637	0.613
PC5	0.710	0.699	0.736	0.751
tomcat	0.563	0.533	0.566	0.590
Mean	0.620	0.605	0.620	0.625
Total shaded	3	1	3	4

Table 9. The obtained results of T-BWOA in terms of error rate (ER)(%)

Dataset	T1-BWOA	T2-BWOA	T3-BWOA	T4-BWOA
Cm1	0.148	0.151	0.16	0.159
JM1	0.042	0.042	0.044	0.042
KC1	0.154	0.150	0.158	0.158
KC3	0.210	0.219	0.214	0.202
MW1	0.104	0.140	0.100	0.112
PC1	0.087	0.098	0.099	0.116
PC2	0.012	0.010	0.011	0.012
PC3	0.147	0.132	0.148	0.139
PC4	0.138	0.157	0.131	0.133
PC5	0.027	0.026	0.024	0.023
tomcat	0.121	0.108	0.100	0.11
Mean	0.1081	0.1120	0.1080	0.1096
Total shaded	3	4	3	3

Table 10. The averaged selected features obtained by the proposed method (T-BWOA-KNN)

Datasets	T1-BWOA-KNN	T2-BWOA-KNN	T3-BWOA-KNN	T4-BWOA-KNN
CM1	3	2	4	4
JM1	2	2	2	2
KC1	7	4	7	6
KC3	4	4	6	4
MW1	5	6	5	6
PC1	5	3	4	4
PC2	2	2	3	2
PC3	5	2	7	6
PC4	8	5	7	7
PC5	3	2	4	3
TOMCAT	5	3	5	5
Total shaded	4	9	2	3

Among the six comparatives above, four cases clarified that T2-BWOA had produced the best performance for solving feature selection in SDP compared with the others.

For more proof of the effectiveness of the suggested method (T-BWOA_KNN) in selecting the optimal features that yield the maximum classification accuracy, table 10 shows the number of best features for each method on average. The smaller number of features means that the performance of the method is better. As shown in table 10, T2-BWOA has fewer features for nine datasets from eleven compared to other methods. For example, for CM1 dataset, T2-BWOA selected around two features compared to three, four, four, and four features for T1-BWOA-KN, T3-BWOA-KNN, and T4-BWOA-KNN, respectively. Thus, the proposed T2-BWOA has the minimum number of the selected features with the most of the datasets by comparing it with other methods. So, it is considered the best method compared with others for solving feature selection problem in SDP.

For more clarity of the experiment results, a boxplot is used for all datasets in terms of accuracy to show the minimum, median, and maximum accuracy of the obtained results for running the experiments ten times autonomously where the red line for each block of the dataset represents the median. As shown in Fig. 3, the T2-BWOA-KNN method achieved the highest values for the PC2 dataset. On the other hand, it achieved the worst values for the KC3 dataset. At the same time, it produces the best performance in terms of accuracy for four datasets from eleven compared to other methods that produced the best accuracy for three datasets only.

5. Comparison of T-BWOA-KNN with other methods from the literature

After determining that T-BWOA-KNN has produced the best results for solving feature algorithms in SDP in terms of accuracy and minimum selected features based on the aforementioned results, the performance of T-BWOA-KNN has been compared with the performance of other researches from the literature that experimented on the same datasets in terms of accuracy and selected features.

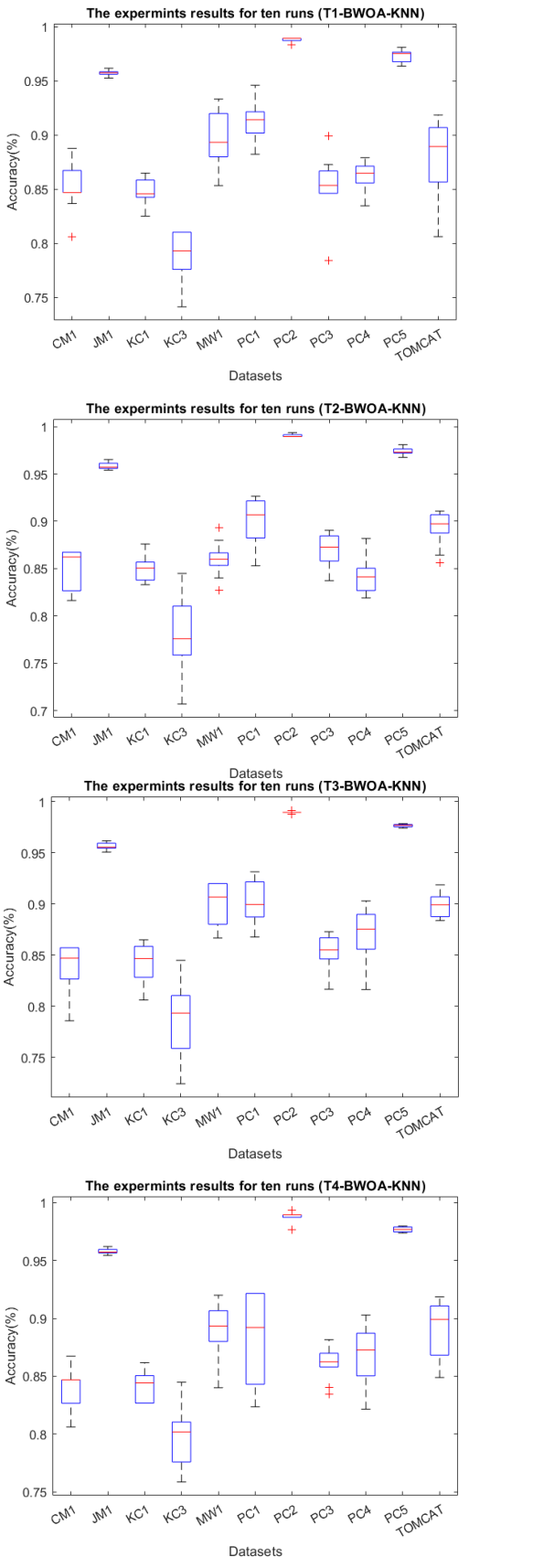


Fig. 3. Boxplot of all datasets for the proposed method(T1~T4-BWOA-KNN) in terms of accuracy

The first research enhanced wrapper feature selection (EWFS) based on a dynamic re-ranking strategy by deploying two classifiers: Decision Tree (DT) and Naïve Bayes (NB) [5]; the second research used correlation-based feature subset selection (CFS) with metaheuristic algorithms such as GA, BAT, PSO, FS, and AS with multiple classifiers such as KNN [3]; the third research used rank aggregation which depends

on feature selection method with multifilters which is called (RMFFS) with two classifiers which are Decision Tree (DT) and Naïve Bayes (NB) [4]; the last research was proposing a hybrid method by using the multi-layer perceptron (MLP) with multi-filter feature selection technique called (MLP-FS) [16]. The comparison results are shown in table 11 and table 12, where/' means that the value for this dataset is not available and the Shaded cells refer to the best-obtained results. It is clear that the performance of T-BWOA-KNN was better than other methods in most datasets in terms of accuracy and the number of selected features. For instance, the result in dataset PC5, an improvement of at least more than 30% in terms of selected features and 21% in terms of accuracy can be achieved with T-BWOA-KNN as shown in table 11 and table 12, respectively.

In summary, the obtained results support the objective of this research that the taper-shaped transfer function can improve the overall performance of WOA with KNN classifiers. Last but not least, it can be concluded that T-BWOA-KNN is a useful tool for solving feature selection problems in SDP.

For more investigation and in order to show the significance of the proposed method (T-BWOA-KNN), the null hypothesis (H0) and alternative hypothesis (H1) are constructed as shown below:

H0: there is no significant difference between the accuracy of the proposed method (T-BWOA-KNN) with other methods from the literature.

H1: there is a significant difference between the accuracy of the proposed method (T-BWOA-KNN) with other methods from the literature. Thus, the Kendall W test has been implemented to show the significance and the ranking of the proposed method (T-BWOA-KNN) with other methods in the state-of-the-art that are used KNN classifier only in terms of the accuracy for mutual datasets (i.e. CM1, KC1, KC3, MW1) using Kendall W test results. The results of the Kendall W test for the mutual datasets indicate that the P-value (0.002) of the test is lower than α (0.05) which means that the condition of the null hypothesis, (H0) is rejected and the alternative hypothesis (H1) is accepted. This indicates that the proposed method (T-BWOA-KNN) is significant.

As shown in table 13, the performance of the proposed model (T-BWOA-KNN) is more significant than other methods in terms of accuracy for solving SDP based on p-value. Also, the overall ranking of all methods indicates that the proposed (T-BWOA-KNN) has the top ranks among other methods that used KNN classifier, as demonstrated in figure 4.

Table 11. Comparison of T-BWOA-KNN to other methods in the state of art in terms of selected features

Dataset	The Proposed method	[5]		[13]									
	T2-BWOA-KNN	NB+EWFS	DT+EWFS	CFS+KNN+GA	CFS+KNN+BAT	CFS+KNN+PSO	CFS+KNN+FS	CFS+KNN+AS	CNS+KNN+GA	CNS+KNN+BAT	CNS+KNN+PSO	CNS+KNN+FS	CNS+KNN+AS
CM1	2	4	7	7	5	8	7	5	12	12	6	15	8
JM1	2	/	/	/	/	/	/	/	/	/	/	/	/
KC1	4	2	4	8	4	8	4	2	11	17	16	16	17
KC3	4	3	3	2	2	3	3	2	9	17	6	12	13
MW1	6	3	3	8	9	7	9	7	11	17	8	17	13
PC1	3	5	6	/	/	/	/	/	/	/	/	/	/
PC2	2	/	/	5	5	5	5	6	15	17	9	17	16
PC3	2	3	5	/	/	/	/	/	/	/	/	/	/
PC4	5	6	3	/	/	/	/	/	/	/	/	/	/
PC5	2	3	6	/	/	/	/	/	/	/	/	/	/
TOMCAT	3	6	4	/	/	/	/	/	/	/	/	/	/

Table 12. Comparison of T-BWOA-KNN to other methods in the state of art in terms of accuracy

Dataset	proposed	[5]		[13]					[29]		[14]
	T2+BWOA+KNN	NB+EWFS	DT+EWFS	CFS+KNN+GA	CFS+KNN_+BAT	CFS+KNN_+PSO	CFS+KNN_+FS	CFS+KNN+AS	NB+RMFFS	DT+RMFFS	MLP+FS
CM1	84.9	87.16	86.63	77.68	77.37	80.43	81.04	78.59	73.3	61.8	89.795
JM1	95.8	/	/	/	/	/	/	/	/	/	80.44
KC1	85.0	75.3	75.9	71.26	72.98	70.57	70.40	71.69	78.2	65.1	77.6504
KC3	78.1	82.47	85.41	78.87	78.87	75.77	74.23	75.26	71.0	68.1	82.758
MW1	86.0	90.0	89.2	84	84.4	84	84	82	/	/	92.000
PC1	90.2	91.9	91.9	/	/	/	/	/	/	/	96.078
PC2	99.0	/	/	96.54	96.12	96.81	95.84	95.29	/	/	97.695
PC3	86.8	84.59	86.54	/	/	/	/	/	79.8	66.6	85.126
PC4	84.3	82.67	88.89	/	/	/	/	/	/	/	88.97
PC5	97.4	74.87	76.04	/	/	/	/	/	/	/	74.803
TOMCAT	89.2	90.96	92.02	/	/	/	/	/	/	/	/

Table 13. Ranking of the proposed method (T-BWOA-KNN) with other methods using Kendall W test

Mutual Datasets	W	P	Rank methods	
CM1, KC1, KC3, MW1	0.745	0.002	T1_BWOA_KNN	8.25
			T3_BWOA_KNN	7.25
			T4_BWOA_KNN	7.25
			T2_BWOA_KNN	6.75
			CFS_KNN_BAT	4.12
			CFS_KNN_GA	3.38
			CFS_KNN_PSO	3
			CFS_KNN_FS	2.5
			CFS_KNN_AS	2.5

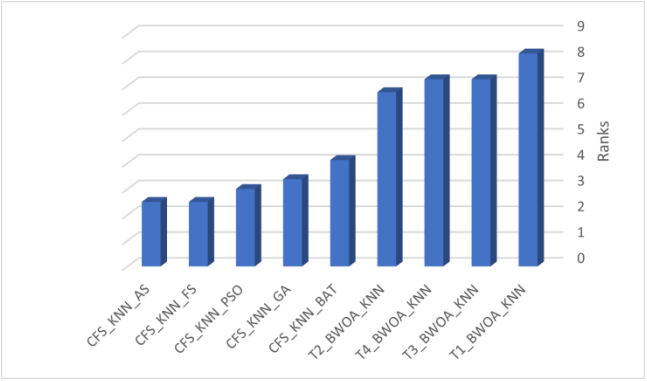


Fig. 4. The ranking of (T-BWOA-KNN) with other methods from the literature

6. Conclusion

This paper includes evolving the binary whale optimization algorithm (BWOA) by using the taper shaped transfer function to convert the continuous search space to binary search space for solving feature selection problems in SDP.

The main purpose of this research was to select the minimum number of relevant selected features with the highest accuracy for solving SDP.

The proposed method (T-BWOA-KNN) has been applied on eleven datasets which are CM1, JM1, KC1, KC3, MW1, PC1, PC2, PC3, PC4, PC5, TOMCAT that are obtained from NASA and promise repository. These datasets are varies based on the number of projects (patterns), attributes (features), and defect ratio. Each experiment has been used KNN classifier and repeated ten times autonomously to show the performance of the proposed method (T-BWOA-KNN) based on seven evaluations metrics.

The experimental results have shown that the proposed method T-BWOA-KNN has produced the highest classification accuracy and the minimum number of the selected features for most of the datasets compared to other methods. Also, it was showed that the performance of feature selection methods depends on dataset and the used classifier. In addition, the proposed method (T-BWOA-KNN) has the top ranks among other methods from the literature that used the mutual datasets and KNN classifier in terms of accuracy. As future works, utilizing another classifier such as a multilayer perceptron (MLP) or support vector machine (SVM) instead of the KNN. Also, suggest a new feature selection technique by hybridization Quasi-Oppositional Method with binary whale optimization algorithms as a feature selection method in SDP.

References

[1] Adamu A. et al.: An hybrid particle swarm optimization with crow search algorithm for feature selection. Machine Learning with Applications 6, 2021, 100108.
[2] Al Qasem O., Akour M.: Software fault prediction using deep learning algorithms. International Journal of Open Source Software and Processes (IJOSSP) 10(4), 2019, 1–19.
[3] Balogun A. O. et al.: Performance analysis of feature selection methods in software defect prediction: a search method approach. Applied Sciences 9(13), 2019, 2764.
[4] Balogun A. O. et al.: Rank aggregation-based multi-filter feature selection method for software defect prediction. Advances in Cyber Security: Second International Conference – ACeS 2020, 2021.
[5] Balogun A. O. et al.: Software defect prediction using wrapper feature selection based on dynamic re-ranking strategy. Symmetry 13(11), 2021, 2166.
[6] Balogun A. O. et al.: An adaptive rank aggregation-based ensemble multi-filter feature selection method in software defect prediction. Entropy 23(10), 2021, 1274.
[7] De Souza R. C. T. et al.: A V-shaped binary crow search algorithm for feature selection. IEEE Congress on Evolutionary Computation – CEC, 2018.
[8] Fan G. et al.: Software defect prediction via attention-based recurrent neural network. Scientific Programming 2019, 6230953.
[9] Gad A. G. et al.: An improved binary sparrow search algorithm for feature selection in data classification. Neural Computing and Applications 34(18), 2022, 15705–15752.
[10] Hamed A. et al.: Algamal, Improving binary crow search algorithm for feature selection. Journal of Intelligent Systems 32(1), 2023, 20220228.
[11] Hassouneh Y. et al.: Boosted whale optimization algorithm with natural selection operators for software fault prediction. IEEE Access 9, 2021, 14239–14258.
[12] He Y. et al.: Novel binary differential evolution algorithm based on Taper-shaped transfer functions for binary optimization problems. Swarm and Evolutionary Computation 69, 2022, 101022.
[13] Hossin M., Sulaiman M. N.: A review on evaluation metrics for data classify-cation evaluations. International journal of data mining & knowledge management process 5(2), 2015.
[14] Hussien A. G. et al.: Binary whale optimization algorithm for dimensionality reduction. Mathematics 8(10), 2020, 1821.
[15] Hussien A. G. et al.: S-shaped binary whale optimization algorithm for feature selection. Recent Trends in Signal and Image Processing: ISSIP 2017.
[16] Iqbal A., Aftab S.: A Classification Framework for Software Defect Prediction Using Multi-filter Feature Selection Technique and MLP. International Journal of Modern Education & Computer Science 12(1), 2020.
[17] Jureczko M., Madeyski L.: Towards identifying software project clusters with regard to defect prediction. Proceedings of the 6th international conference on predictive models in software engineering. 2010.
[18] Landis J. R., Koch G. G.: The measurement of observer agreement for categorical data biometrics 1977, 159–174.
[19] Legendre P.: Species associations: the Kendall coefficient of concordance revisited. Journal of agricultural, biological, and environmental statistics 10, 2005, 226–245.
[20] Mirjalili S., Lewis A.: The whale optimization algorithm. Advances in engineering software 95, 2016, 51–67.
[21] Rathore S. S., Kumar S.: A decision tree logic based recommendation system to select software fault prediction techniques. Computing 99, 2017, 255–285.
[22] Shepperd M. et al.: Data quality: Some comments on the nasa software defect datasets. IEEE Transactions on software engineering 39(9), 2013, 1208–1215.
[23] Shepperd M. et al.: Nasa mdp software defects data sets. Figshare Collection, 2018.
[24] Tohka J., Van Gils M.: Evaluation of machine learning algorithms for health and wellness applications: A tutorial. Computers in Biology and Medicine 132, 2021, 104324.
[25] Tumar I. et al.: Enhanced binary moth flame optimization as a feature selection algorithm to predict software fault prediction. IEEE Access 8, 2020, 8041–8055.
[26] Turabieh H., Mafarja M., Li X.: Iterated feature selection algorithms with layered recurrent neural network for software fault prediction. Expert systems with applications 122, 2019, 27–42.
[27] Witten I. H., Frank E., Hall M. A.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc. 2011.
[28] Xu Z. et al.: Software defect prediction based on kernel PCA and weighted extreme learning machine. Information and Software Technology 106, 2019, 182–200.
[29] Zhou T. et al.: Improving defect prediction with deep forest. Information and Software Technology 114, 2019, 204–216.
[30] Zhu K. et al.: Software defect prediction based on enhanced metaheuristic feature selection optimization and a hybrid deep neural network. Journal of Systems and Software 180, 2021, 111026.

M.Sc. Zakaria A. Hamed Alnaish
e-mail: zakriahamoalnaish@uomosul.edu.iq

He is master in computer science from university technology Malaysia, faculty of computing and got best student award. Currently, he works as lecturer in College of Science/University of Mosul and Ph.D. candidate in Computer Science Department, College of Computer Science and Mathematics, University of Mosul.
Research interest: machine learning, fuzzy logic, optimization algorithms, software engineering.

http://orcid.org/0000-0002-7597-5326

Prof. Dr. Safwan O. Hasoon
e-mail: Dr.safwan@uomosul.edu.iq

He is a doctor and full professor in artificial intelligence. He works College of Computer Science and Mathematics, University of Mosul.
Research interest: machine learning, artificial intelligence, artificial neural network.

http://orcid.org/0000-0002-3653-3568

http://doi.org/10.35784/iapgos.4261 *received: 16.07.2023 | revised: 19.08.2023 | accepted: 22.10.2023 | available online: 20.12.2023*

USE OF THE CDE ENVIRONMENT IN TEAM COLLABORATION IN BIM

Andrzej Szymon Borkowski, Jakub Brożyna, Joanna Litwin, Weronika Rączka, Aleksandra Szponarowicz
Warsaw University of Technology, Faculty of Geodesy and Cartography, Warsaw, Poland

Abstract. *In project processes, group collaboration and project documentation management are important aspects. In order for the cooperation of all project participants to be effective, it should be based first and foremost on adequate and effective communication. All project participants should use such solutions so that they can exchange, manage and combine information quickly and efficiently throughout the entire investment process, thus providing a complete picture of the situation. To this end, it is necessary to develop a catalogue of good practices supported by a variety of examples, as well as rules for group cooperation when using a CDE-type solution. The aim of this article was to show the advantages and benefits as well as the disadvantages and limitations in group collaboration when working on a single BIM model.*

Keywords: BIM, building information modelling, CDE, common data environment, team collaboration

WYKORZYSTANIE PLATFORMY CDE WE WSPÓŁPRACY ZESPOŁOWEJ W BIM

Streszczenie. *W procesach projektowych ważny aspekt stanowi współpraca grupowa oraz zarządzanie dokumentacją projektową. Aby współpraca wszystkich uczestników projektu była efektywna, powinna opierać się przede wszystkim na odpowiedniej i efektywnej komunikacji. Wszyscy uczestnicy projektu powinni używać takich rozwiązań, aby przez cały proces inwestycyjny mogli szybko i sprawnie wymieniać się informacjami, zarządzać nimi i łączyć je ze sobą, dostarczając w ten sposób kompletny obraz danej sytuacji. W tym celu konieczne jest opracowanie katalogu dobrych praktyk podpartych różnorodnymi przykładami oraz zasad współpracy grupowej przy stosowaniu rozwiązań typu CDE. Celem artykułu było pokazanie zalet i korzyści oraz wad i ograniczeń we współpracy grupowej podczas pracy na jednym modelu BIM.*

Słowa kluczowe: BIM, modelowanie informacji o budynku, CDE, wspólne środowisko danych, współpraca grupowa

Introduction

BIM (Building Information Modelling) is no longer new to the AEC (Architecture, Engineering, Construction) industry. With the rapid development of BIM, it has become apparent that the problem is not the use of BIM software that can be mastered, but collaboration between designers or trades [12]. Researchers and computer scientists have developed virtual and physical collaboration platforms for AEC and FM (Facility Management). Improved communication through the use of wireless and cloud technologies has been identified as an important factor in maintaining the quality of this collaboration. Interoperability and online storage of models is already widely used in BIM [11]. Holzer, in his 2014 reflections, notes that practitioners are well aware that the quality of BIM itself depends on a deeper understanding of the building construction process. There is a danger associated with the use of BIM by inexperienced designers, which can imply inefficiencies in the design process as well as in project delivery [4]. In addition to familiarity with BIM technology, it is crucial to develop critical thinking skills, openness and attentiveness to innovation, the ability to adapt to rapidly changing environments and to communicate and manage information efficiently within a multidisciplinary team. BIM, with the development of the software, has become a symbol of collaboration, but the adversarial nature of corporate branding and market dominance has led to a plethora of mutually incompatible BIM offering [18].

Let the multitude of problems (integration with GIS (Geographic Information System), macro to micro modelling of objects, insufficient interoperability, etc.) faced by researchers, computer scientists and practitioners [15] attest to how difficult group collaboration is. Designers use various workarounds to solve the problems because BIM is evolving all the time and the development of technology has not kept pace with the development of the methodology [5]. At this point, it is important to clarify what BIM is and what it is not. BIM is not software. BIM is not Revit, as some people claim. The BIM process is not about building stunning 3D models that will ultimately culminate in photorealistic exciting visualizations. BIM is a 'process' that completely changes the status quo in the construction industry. A new paradigm is emerging: BIM is a new way of organizing work throughout the construction project process. A BIM model is a collection of knowledge and information about a construction object that forms the basis for decision-making throughout the life cycle of that object. Eastman and his group define it as follows: Building Information Modelling (BIM) is a collaborative way of storing, sharing, exchanging and managing multidisciplinary information throughout the lifecycle of a construction project, including the planning, design, construction, operation, maintenance and demolition phases [3]. Eastman has repeatedly emphasized that BIM is about activity undertaken by people, not a model built by people. BIM is variously defined and understood by both established organizations and many researchers. According to the authors, BIM can be considered from two perspectives – a broader one and a narrower one. BIM *sensu largo* is a process based on the collaboration of people, IT systems, databases and software. In a much broader perspective, it can also include hardware, tangible and intangible resources or knowledge. BIM *sensu stricto* is the relational database of a building object accompanying it throughout its life cycle.

Project work in a group should be based on communication – unfortunately equipment often fails or problems arise at other levels. Many times a single channel of communication is not enough to exchange information - it is important to bear in mind that reporting problems and trying to resolve them can be done using traditional communication channels e.g. email, chat, instant messaging or telephone contact. However, in a mature level two BIM approach (according to the Bew-Richards wedge), one works using the CDE (Common Data Environment) platform, which should be the only communication channel [17].

1. Structure of the Common Data Environment

The CDE platform is a tool that specifically influences the collaboration of all those involved in a project. It acts as a virtual place where the exchange of information about a construction project takes place. They include not only the BIM model and the elements created in this environment, but also documentation, graphic resources and various components collected during the design process [9]. Using one common source of information and updating it systematically improves collaboration and reduces errors, and duplication. This is crucial for project delivery, collaboration efficiency, model hygiene and work culture. Conduct within the use of the BIM environment is structured by the ISO 19650 standard, which consists of six parts (four of which are published). It defines a unified framework for the collaborative and effective management of a facility throughout its life cycle. Importantly, there is freedom to create your own requirements as to ISO 19650 and CDE,

an example being the Victorian Digital Asset Strategy (VDAS) in Victoria [10].

The operation of the CDE platform is concentrated in the four working areas [8] shown in the diagram (figure 1).

From the point of view of the owner and the implementation of a CDE project, the control functions are strengthened and this ultimately has the effect of reducing costs, increasing quality (while maintaining all necessary parameters) and reducing time delays [12]. The CDE environment allows information to be collected from a wide range of industries involved in the design, implementation, and subsequent use, these include architectural, structural, building, infrastructure, landscape, electrical or plumbing data.

One example of multi-discipline collaboration is the Pan Borneo Highway project, which connects two states: Malaysia, Sabah and Sarawak, with Brunei and the Kalimantan Region in Indonesia [1]. In the course of this implementation, it was crucial to provide a suitable workspace for engineers or entrepreneurs dispersed around the world. This meant, among other things, integrating various data, files and knowledge, as well as optimizing costs and working times, which were important for such a large investment Implementing a CDE for infrastructure projects provides a measurable return on investment (ROI), such as saving time for searching, validation and accessing project information.

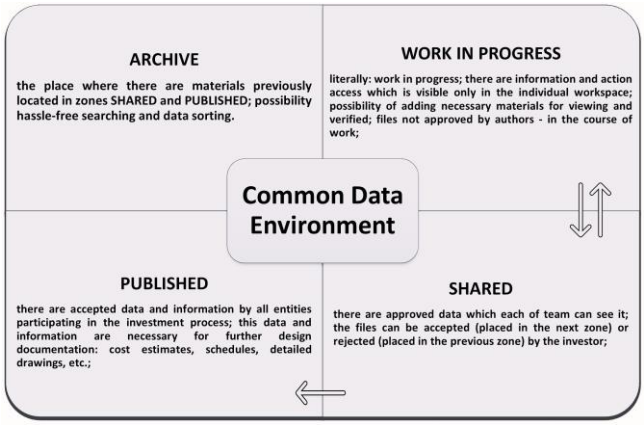


Fig. 1. Diagram illustrating the functioning of the data sharing environment (CDE) structure

2. The practical implementation of CDE

The CDE platform plays a key role in the management of information and documentation in BIM-based projects, which contributes to a more efficient flow of data between project participants. In practice, CDE platforms, such as Bentley ProjectWise or Autodesk BIM 360, support the delivery of diverse construction and engineering projects by enabling the centralization of data and resources and close collaboration across trades. Group collaboration enables a project team to work continuously using one common model (not necessarily one file). Effective group collaboration is not only about software that allows people to work on the same project, but also about properly defining the standards, stages, goals and ways of this collaboration [6]. Examples of the use of the CDE platform include the Crossrail project (Elizabeth Line) in the UK. However, the CDE platform can also have a more common application, which is its use in smaller-scale projects, such as the development of residential structures or construction projects requiring extensive interprofessional collaboration.

Crossrail (Elizabeth Line) represents one of the largest construction projects in Europe, which aimed to meet London's future transport needs and support the UK's continued economic growth. The project involved the construction of 118 km of rail line, including 21 km through tunnels under central London, leading to Shenfield and Abbey Wood in the east. The complexity of this project required the use of an integrated design environment to enable the project to be delivered. The company behind the Crossrail project has created an intelligent 3D model with linked databases managing the associated design data, meaning the project is largely compliant with the UK GCS BIM Level 2 requirements set out in 2016. The Government Construction Strategy (GCS) is the BIM technology implementation strategy implemented by the UK government. It shows that all government projects drawn up after 2016 require the use of BIM technology at maturity level two, which is based on collaboration between project teams, the sharing of information in both 2D and 3D formats, and the use of a common data environment – the CDE – to manage project documentation. The CDE-based working model has contributed to improved management and increased project quality through the use of the 3D model. The creation of a consistent and complete 'as-built' BIM model has also helped to maximize the benefits, as it has been handed over to the railway operator, who can carry out inspections and maintenance based on it [14].

3. The challenges encountered in collaborative work on a shared model

Group collaboration on a single model is intended as a response to common problems encountered during the 'traditional' design process. Such challenges include: lack of ongoing updating of the model, which results in designing based on outdated data and, consequently, the need to repeat the completed work. Lack of workflow – this is particularly significant when multi-discipline teams are working together, where a minor modification to the design of one trade can make the design of another trade immensely easier or provide the opportunity to introduce a much more efficient solution. Other challenges are the loss of information and the need to recreate it, which can lead to errors and discrepancies, the difficult use of the results of other team members' work, resulting in designers from different branches having to enter the same data multiple times. The last and greatest challenge in group collaboration, reduced by group collaboration on a single model, are problems in the coordination of the model, design, collision or quantity statements, which can carry the uncertainty of generating up-to-date design documentation when they need to be delivered to the construction of drawings while still in the design process [6].

Group work on a single model is intended to improve the efficiency of collaboration between project participants and the quality of the final product. In the context of information management and data exchange within the CDE, it solves one of the major challenges of keeping project documentation consistent and up-to-date. In the case of the implementation of construction projects, such as roads or residential buildings, it is important to adapt the standards and formats of data exchange to the specifics of the local construction market, which may require government intervention to standardize this environment. For example, in the Czech Republic, where road projects are subject to a complex legislative system, frequently changing regulations and a complex decision-making structure of local authorities, a properly defined CDE environment can contribute to making projects more resilient to changes in the external environment [13].

4. Georeferencing of the BIM model

The correct positioning of the BIM model in GIS space enables reliable analyses and simulations. The integration of BIM and GIS is an important technological and scientific trend [2]. In the near future, it can be seen as one of the biggest challenges for the construction industry, in order to use the vast amount of data more efficiently and achieve synergies. Integration can be considered at the application, process and data level. From the project team's point of view, integration at the application level is the easiest. Typically, users of BIM modelling software look for specific functionality in the application that allows them

to import GIS data. Unfortunately, in many cases such functionalities are either not present at all or are limited to only loading web map services (Web Map Service).

Many popular BIM software packages have certain limitations when working with GIS data. These limitations often affect or prevent the georeferencing of BIM models. The Autodesk Revit application has a maximum distance limit from its internal datum of 10 miles (over 16 kilometers). Once this limit is exceeded, the graphical representation of the components used becomes less reliable and less accurate. This limitation applies to geometry directly created in Autodesk Revit, as well as geometry derived from an import or link. As databases, city models and planning tools are developed in GIS-enabled software such as ArcGIS or QGIS, they are often geographically referenced and located very far from the origin of the BIM models. Directly importing this information, or entering the coordinates manually into Autodesk Revit, often results in a 'Geometry in file is greater than 10 miles (16 km)' error. So the workspace is limited in both directions – a total of 20 miles (more than 32 km), and the designer must otherwise establish the coordinates of the survey point (x,y,z) and give the appropriate GIS coordinate system.

The .rvt project file with fixed X,Y,Z coordinates of the base point should be loaded in ESRI ArcGIS Pro, which supports both GIS and BIM data. Pre-connecting the ESRI ArcGIS GeoBIM service with Autodesk Construction Cloud Services will provide the ability to directly load BIM data into a local or global scene in GIS. To geo-reference the BIM model, load any GIS reference data in the desired coordinate system (e.g. 2000 lane 7 system), then load the BIM model and move it to the anticipated location. Once the coordinate system is correctly assigned, the .rvt file receives two additional companion files with extensions .prj and .wld3. The .prj file is a text file containing information defining the coordinate system, data and map projection. A projection file with the same name as the BIM or CAD file is used to define the coordinate system of this file (<filename>.prj). The .prj file is sometimes referred to as a universal projection file. The .wld3 file, on the other hand, is often referred to as the 'world file' (world file). Using this file to store control points is a best practice for sharing and reusing links in other projects and map documents. ArcGIS Pro uses the file name and its location to associate the .wld3 file with a specific CAD dataset or BIM model (.rvt or .ifc). This file is similar to a world file, but contains coordinates so that the file can be placed in the correct x,y,z location. Correct geo-referencing of the BIM model enables it to be further used by the designer to produce their model or disseminated more widely e.g. to tradesmen. This then facilitates the federation (together) of industry models into a whole. Team members' access to the CDE provides access to the model files with the correct geo-referencing.

5. Object of research

The subject of the study was a BIM model of a two-story building with an office function (figure 2). The building was surrounded on all sides by paved surfaces in the form of a car park (figure 3). The ground floor comprised rooms such as a storage room, three bathrooms, an office and a conference room. The rooms comprising the first floor included an office, two bathrooms and a rest room. Space was also designated for a lift and a staircase. The first floor was laid out with a corridor providing access to each room. The main intention of the project was to test the CDE platform by converting/adapting the aforementioned office function of the building into residential, while at the same time increasing the biologically active area of the surrounding space. It is worth mentioning that the work was carried out on a previously provided finished model. All modifications and tests of the functioning of the CDE environment were experimental and carried out in a controlled environment.

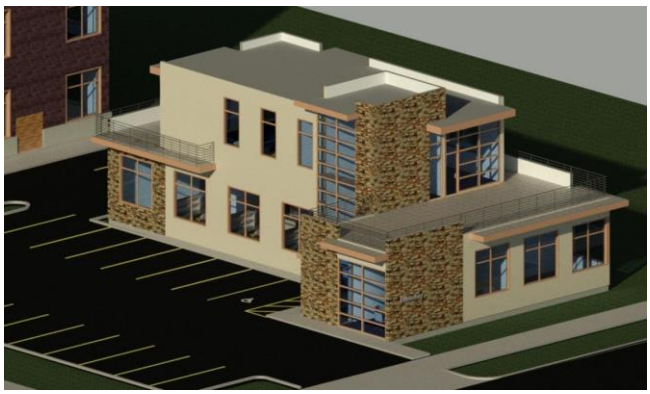


Fig. 2. Office building

Fig. 3. Office building with immediate surroundings

6. Group cooperation on a central model

In the case study presented here, the main focus of the group collaboration on the central model was to change the function of the site from office to residential. To this end, the interior of the building and its surroundings were redesigned to meet the requirements of the newly assigned function. The scope of work included, among other things, changing the site's surface and landscaping, as well as adapting the building by modifying the layout and interior finishes. Two teams of three people worked on the project simultaneously: one team worked on the interior of the building and the other on the exterior. In the CDE environment, each user used a local version of the model, which was a 'branch' of the central model. Therefore, during the process of updating the model, version continuity is maintained and, as a result of combining data from both models, the new version shows the changes made. Such a solution streamlines the work by reducing the need to send the updated files to the other team members themselves and reducing the risk of data inconsistency [16].

Nevertheless, some difficulties were encountered during the collaboration. During the project, each team member worked on a local copy of the central model and uploaded their changes to the central model. However, the synchronization process was not performed automatically in real time, so after each work step, a two-step synchronization had to be performed manually by clicking on two different synchronization icons in the BIM design software and then publishing the changes made. The synchronized model did not immediately appear to the other users. For this reason, they had to update their model first by selecting the icon to reload the latest changes in order to see the effects of the other team members' work (figures 4 and 5). Such asynchronous working has its strengths and weaknesses.

Fig. 4. View of the synchronized model from the user making the changes

Fig. 5. View of the non-synchronized model from the non-synchronized user level

When making changes, it proved convenient for this type of collaboration to be able to use components uploaded to the project by other users. As a result, project consistency was easily maintained without having to send separate family files to each other. In addition, the option to modify changes made to the project by other users also optimized and accelerated the work. This made it possible, for example, to move furniture inserted by another user without having to send an edit request for the object in question.

7. The importance of CDE for group and interprofessional cooperation

The CDE platform is extremely important as it enables collaborative projects involving multiple designers, managers, investors and others involved in a project or development. The simultaneous, coordinated work of many people on a single model increases the efficiency of the BIM process, amplifying the benefits provided by building information modelling technology. Another advantage is the ability to enrich the model with user-added elements such as families. This makes the process of modifying objects more efficient, which, moreover, can also be edited by other users.

However, the use of the platform requires the observance of certain precautions. The CDE system provides a tool for coordinating work and sharing results easily. However, it is not a technology that allows simultaneous, real-time modifications to a single model file by several users. This therefore imposes the need for mutual time synchronization. In addition, a slightly cumbersome aspect of working with the platform is the need to approve changes by double synchronization of local files with the central model each time, resulting from the method of functioning of the Revit application tools designed for this purpose. Some problems with the transparency of the system were also identified, which relate to the system of informing the user about the validity of particular file versions. This is because newly introduced changes are not always marked as new versions of the project.

Fig. 6. Visualization of the external adaptation of the site for new functions

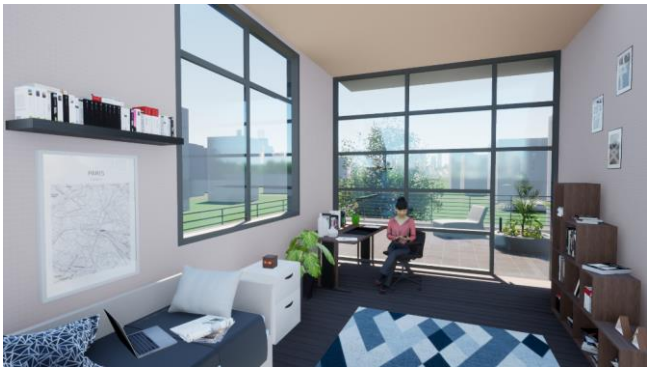


Fig. 7. Visualization of the interior after rearrangement for new functions

The visualizations produced (figures 6 and 7), showing the end result of the group collaboration on the model. By using the CDE platform, it was possible to achieve the main objective, the transformation of the original function of the building to residential. The coordinated collaboration between the different teams provided a practical demonstration of how the platform works. Group and interprofessional collaboration can be fraught, however, with risks due to a lack of communication and adherence to commonly agreed rules. Lack of these factors can result in duplicate content, working on erroneous files or failing to develop common and mutually acceptable solutions. The effectiveness and efficiency of using the CDE platform can be achieved by following basic rules such as: updating uploaded data systematically, using it as the only place to collect all content, from drawings to emails, which should only be corresponded to here. An important aspect is to assign specific roles to all those involved in the design process and give them the appropriate permissions in moving and operating within the CDE environment [9]. In summary: one of the most influential criteria when implementing CDE is data security and, on the other hand, the biggest obstacle to its use is the lack of collaboration between users and the lack of discipline [7].

8. Discussion

The use of the CDE platform in the management of construction and engineering projects brings numerous benefits that streamline the design and collaboration processes between the various project participants. It is a not inconsiderable improvement that represents a step towards completely moving all project phases to BIM technology. The use of the CDE platform meets the definition of BIM maturity at level 2 according to the Bew-Richards wedge (PAS 1192-2:2013). However, despite this huge pool of applications, the use of the CDE platform has both benefits and drawbacks.

In terms of the spatial localization of a project, ArcGIS GeoBIM enables it to be precisely located in a geographic information system (GIS), allowing spatial aspects to be analyzed and taken into account in project planning and implementation.

One of the key strengths of the CDE platform is the ability to modify changes made by other users, enabling dynamic collaboration and efficient management of different versions of documents and models. In addition, the system allows the use of families loaded into the project by other users, which provides access to updated components and speeds up the process of creating models.

Another key aspect of the CDE platform is the interoperability of .IFC files, which allows data to be exchanged between different applications and project management systems, regardless of the software used. This facilitates collaboration between specialists from different disciplines and increases the efficiency of project processes. The ability to view IFC files in free browsers also promotes the efficiency of the project team. In this way, all project participants, even those who do not have access to full versions of specialized tools, can verify and analyze models and adapt their activities to the current project requirements. As a result, the CDE platform provides an environment that fosters effective and flexible collaboration between teams and enhances the quality of projects (figure 8).

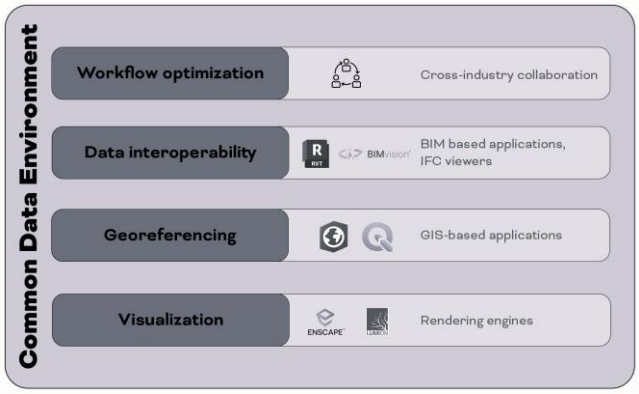


Fig. 8. Diagram showing the functionalities of the CDE platform

Despite the numerous advantages of the CDE platform, there are also some drawbacks that can affect the efficiency of collaboration and project delivery times. First and foremost, the need to work twice (message when synchronizing local files with the central model in Revit can be cumbersome and time-consuming for users. In addition, the software does not always recognize changes made as a new version of the project, which can lead to ambiguity in marking and tracking progress. The insufficient precision of the project location, given by address, compared to geographical coordinates, can make it difficult to accurately represent the project on the map in the Project Home tab. Collaboration in CDE can also be limited by the inability of several people to make changes at the same time, which affects the dynamics of project processes. In addition, the displayed changes only become visible after the project is closed and reopened, which can contribute to a delay in communication between project participants. Furthermore, when performing visualization with the Datasmith add-on in Twinmotion and visualization with Enscape, users may encounter problems with downloading and working on a model that is not available, as well as simplified solids of certain types of objects such as trees.

In terms of working with the .IFC format, familiarity with translators and IFC model building is required, which can be a barrier for some users. Standard import and export settings can lead to problems opening files, and the results of conversion to IFC can be asymmetrical in Archicad and Revit, causing unresolvable problems in one of these applications. Finally, the large file size of the .IFC format can be a limitation on the performance of systems and the transfer of data between users.

9. Conclusion

In summary, the CDE platform brings significant value to the management of construction and engineering projects, streamlining design processes and collaboration between participants. However, despite its numerous advantages, there are some limitations associated with the use of the platform, such as inefficient synchronization, ambiguity of designations or knowledge requirements on open data exchange standards. Nevertheless, the use of CDE is an important step towards the full application of BIM and the potential reduction of the risk of errors through more effective collaboration. In order to take full advantage of the CDE platform, its limitations must be taken into account and collaboration methods must be systematically improved to optimize the efficiency of construction and engineering projects.

References

[1] Akob Z. et al.: Coordination and Collaboration of Information for Pan Borneo Highway (Sarawak) via Common Data Environment (CDE). IOP Conference Series: Materials Science and Engineering, 2019, 12001.
[2] Borkowski A. Sz. et al.: Przegląd dotychczasowych rozwiązań na poziomie aplikacyjnym w zakresie integracji technologii BIM i GIS. Builder 305(12), 2022, 64–69.
[3] Eastman C. M. et al.: BIM handbook: A guide to building information modelling for owners, managers, designers, engineers and contractors. John Wiley & Sons, 2011.
[4] Holzer D.: BIM and parametric design as game changers. 19th International Conference on Computer Aided Architectural Design Research in Asia CAADRIA 2014, 379–388. 2014.
[5] Holzer D.: The BIM Manager's Handbook: Guidance for Professionals in Architecture, Engineering, and Construction. John Wiley & Sons, 2016.
[6] Kasznia D. et al.: BIM w praktyce – standardy – wdrożenie – case study. Wydawnictwo Naukowe PWN, Warszawa 2018.
[7] Lestari S. et al.: Application of common data environment (CDE) as a method of design review in construction project. Journal of Engineering Design and Technology 22(2), 2022, 103–109.
[8] Losev K. Yu.: The common data environment features from the building life cycle perspective. IOP Conference Series: Materials Science and Engineering, 2020, 42012.
[9] McPartland R.: What is the Common Data Environment (CDE)? Cooperate Website NBS for Specifiers, 2016 [https://www.thenbs.com/knowledge/what-is-the-common-data-environment-cde?fbclid=IwAR0jAXE6wF3pZ2PxKZ_3ct-oeZXBXOFeBOOFe7rUdsr1VmAhZXaH4nhnRD8].
[10] McPherson M.: The Ultimate Guide to the Common Data Environment (CDE) in 2023. 12d Synergy, 2022 [https://www.12dsynergy.com/common-data-environment-guide/?fbclid=IwAR0jyvJKXqkhubA22fvwI6wGFoVd3qpdVdaS-eIgpAePIVRAGFBie84bNB0].
[11] Park J. H., Nagakura T.: A Thousand BIM: A Rapid Value-Simulation Approach to Developing a BIM Tool for Supporting Collaboration during Schematic Design. International Journal of Architectural Computing 12(1), 2014, 47–60 [http://doi.org/10.1260/1478-0771.12.1.47].
[12] Race S.: BIM Demystified. RIBA Publishing, 2013.
[13] Radl J., Kaiser J.: Benefits of Implementation of Common Data Environment (CDE) into Construction Projects. IOP Conference Series: Materials Science and Engineering 471(2), 2019.
[14] Smith S.: Building information modelling – moving Crossrail, UK, forward. Management, Procurement and Law 167(3), 2014, 141–151 [http://doi.org/10.1680/mpal.13.00024].
[15] Song Y. et al.: Trends and Opportunities of BIM-GIS Integration in the Architecture, Engineering and Construction Industry: A Review from a Spatio-Temporal Statistical Perspective. ISPRS International Journal of Geo-Information 6(12), 2017, 397 [http://doi.org/10.3390/ijgi6120397].
[16] Tao X., et al.: Smart contract swarm and multi-branch structure for secure and efficient BIM versioning in blockchain-aided common data environment. Computers in Industry 149, 2023, 1–2 [http://doi.org/10.1016/j.compind.2023.103922].
[17] Vlasák P., Čerbák B.: BIM, structural analysis and communication using common data environment (CDE) in the field of water management. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2019, 93–94.
[18] Wierzbicki M. et al.: BIM - history and trends. CONVR2011– International Conference on Construction Applications of Virtual Reality, 2011.
[19] PAS 1192-2:2013; Specification for information management for the capital/delivery phase of construction projects using building information modelling; The British Standards Institution 2013, vii-viii.

Ph.D. Eng. Andrzej Szymon Borkowski
e-mail: Andrzej.borkowski@pw.edu.pl

Graduate of the Faculty of Geodesy and Cartography (Spatial Management) at the Warsaw University of Technology. Since 2019, he has been employed as an assistant professor. He specializes in BIM technology and its holistic use in urban planning. He conducts empirical and theoretical research in BIM applications in engineering design. Author of two books, co-author of 3 monographs and dozens of scientific articles.

http://orcid.org/0000-0002-7013-670X

Eng. Jakub Brożyna
e-mail: jakub.brozyna2@gmail.com

Graduate of Bachelor of Science in Engineering (diploma with honors) at the faculty of Geodesy and Cartography (Spatial Management) at the Warsaw University of Technology. Currently student at Master's program in Spatial Management at Warsaw University of Technology. Co-author of the poster at the 10th Ogólnopolska Konferencja Młodych Badaczy (Jagiellonian University) in Cracow, Poland.

http://orcid.org/0009-0006-7302-1197

Eng. Joanna Litwin
e-mail: asialitwin@onet.eu

Graduate of Bachelor of Science in Engineering in Spatial Management with a specialization in Urban Planning at the Faculty of Geodesy and Cartography, Warsaw University of Technology. Currently a MA student in the same course of study. Holder of a DAAD (German Academic Exchange Service) scholarship.

http://orcid.org/0009-0007-3896-2001

Eng. Weronika Rączka
e-mail: weronika.r@onet.pl

Graduate of the Faculty of Geodesy and Cartography (Spatial Management) at the Warsaw University of Technology. She designs fiber optic networks and she is interested in researching BIM technology in engineering design.

http://orcid.org/0009-0003-7294-6372

Aleksandra Szponarowicz
e-mail: aleksandra.szpnarowicz@gmail.com

A student of the Spatial Planning Bachelor of Engineering Course at Warsaw University of Technology with a specialization in Environmental Conditions in Spatial Planning.

http://orcid.org/0009-0009-2308-1839

http://doi.org/10.35784/iapgos.5586 received: 8.09.2023 | revised: 19.10.2023 | accepted: 14.11.2023 | available online: 20.12.2023

ASYMPTOTICALLY OPTIMAL ALGORITHM FOR PROCESSING SIDE RADIATION SIGNALS FROM MONITOR SCREENS ON LIQUID CRYSTAL STRUCTURES

Dmytro Yevgrafov, Yurii Yaremchuk
Vinnytsia National Technical University, Vinnytsia, Ukraine

Abstract. *An asymptotically optimal compatible algorithm for detecting side radiation signals from the monitor screen on liquid crystal structures and estimating the duration of image immutability on the monitor screen is found, which will better intercept information from monitor screens. The structure of a special technical intelligence tool is justified as a maximum likelihood algorithm for a finite number of unknown quadrature amplitudes of the information leakage signal from the monitor screen on liquid crystal structures.*

Keywords: side electromagnetic radiation and guidance, video card signals, liquid crystal structures, spectra of signals

ASYMPTOTYCZNIE OPTYMALNY ALGORYTM PRZETWARZANIA SYGNAŁÓW PROMIENIOWANIA BOCZNEGO Z EKRANÓW MONITORÓW LCD

Streszczenie*. Znaleziono asymptotycznie optymalny wspólny algorytm wykrywania sygnałów promieniowania bocznego z ekranu monitora na strukturach ciekłokrystalicznych i szacowania czasu trwania niezmienności obrazu na ekranie monitora, który w najlepszy sposób przechwyci informacje z ekranów monitorów. Struktura specjalnego narzędzia wywiadu technicznego jest uzasadniona jako algorytm maksymalnego prawdopodobieństwa dla skończonej liczby nieznanych kwadraturowych amplitud sygnału wycieku informacji z ekranu monitora na strukturach ciekłokrystalicznych.*

Słowa kluczowe: boczne promieniowanie elektromagnetyczne i przesłuchy, sygnały z kart graficznych, struktury ciekłokrystaliczne, widma sygnałów

Introduction

Eighteen years have passed since the publication of Markus Kuhn's work, in which he first investigated the leakage of information from monitors on liquid crystal structures (LCS) [5]. This time was used by developed industrial countries to create special technical means of intelligence (STMI) capable of intercepting information at distances of tens to hundreds of meters [3, 6–8]. At the same time, most open publications on this topic are based on the desire to immediately describe devices for intercepting information, the leakage of which is caused by indirect electromagnetic radiation and interference (IEMR&I), without justifying their structure using the classical theory of optimal signal filtering.

The theory of optimal signal reception against the background of internal receiver noise states that the best signal detection is the one that is most consistent with the IEMR&I signal [9]. Under conditions of a priori uncertainty with respect to most parameters of the leakage signal, a similar signal receiver is an asymptotically Bayesian IEMR&I signal detector – a receiver, which approaches the optimal signal-to-noise ratio at the output of its linear part [10]. Such receivers use maximum likelihood algorithms (MLA), the essence of which is that unknown parameters of the received signal replace them with the most plausible estimates, in the case when the number of unknown parameters is finite.

1. Statement of the problem

We will set a goal to justify the structure of the STMI that implement the MLA. Let the image of the monitor screen on LCS of the computational facilities remain unchanged during the analysis time $T_a \in [T_{a\min}, T_{a\max}]$, $T_{a\min}, T_{a\max}$ – lower and upper limits of the analysis time. Then at the time interval $t \in [0, T_a]$, $T_a >> T_f$, T_f – the period of the monitor scan frames, the signal realisation is analysed

$$x(t) = \begin{cases} n(t), & \text{when the signal is not present} \\ n(t) + s(t, \mathbf{v}_0), & \text{when the signal is present} \end{cases} \quad (1)$$

where $n(t)$ – Gaussian process with zero mean and correlation function

$$B(t,\tau) = M\left[n(t)n(t+\tau)\right] \quad (2)$$

$\mathbf{v}_0$ – vector of actual values of IEMR&I signal parameters $s(t, \mathbf{v}_0)$. It is assumed that $s(t, \mathbf{v}_0)$ is a known function of time t and a vector n with – parameters $\mathbf{v}_0 \in \boldsymbol{\Theta}$, $\boldsymbol{\Theta}$ – parameter space, in which the intervals of each parameter are finite $\Theta_i \in [\Theta_{i1}, \Theta_{i2}], i = 1...n$, or $\boldsymbol{\Theta} \in [\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2]$; $\boldsymbol{\Theta}_1$ – vector of lower parameter values; $\boldsymbol{\Theta}_2$ – vector of upper parameter values.

The decision on the presence of a signal $s(t, \mathbf{v}_0)$ is made when the likelihood ratio function (LRF) is the absolute maximum

$$\sup_{\mathbf{v}\in\boldsymbol{\Theta}} \left[l(T_a, \mathbf{v}) \right] \geq h \quad (3)$$

and about the fact that there is no signal – when $\sup_{\mathbf{v}\in\boldsymbol{\Theta}} \left[l(T_a, \mathbf{v}) \right] < h$, where h – a certain detection threshold that depends on the optimality criterion,

$$l(T_a, \mathbf{v}) = \exp\left[L(T_a, \mathbf{v}) \right]$$
$$L(T_a, \mathbf{v}) = \int_0^{T_a} x(t) V(t, \mathbf{v})\, \mathrm{d}t - \frac{1}{2}\int_0^{T_a} s(t, \mathbf{v}) V(t, \mathbf{v})\, \mathrm{d}t \quad (4)$$

a $V(t, \mathbf{v})$ – solution of the integral equation:

$$\int_0^{T_a} B(t,\tau) V(\tau, \mathbf{v})\, \mathrm{d}\tau = s(t, \mathbf{v}), \qquad t \in [0, T_a] \quad (5)$$

Goal: find (4) by solving integral equation (5), and obtain MLA (3) for an arbitrary vector of unknown parameters $\mathbf{v}$ of IEMR&I signal and unknown time of unchanged image on the monitor screen $T_a \in [T_{a\min}, T_{a\max}]$.

2. Solution of the problem

Let us consider a specific signal of information leakage through IEMR&I in the form of a modulating voltage of one of the three colours of the RGB monitor on the LCS – $s(t, \mathbf{v}_0)$ (see Fig. 1).

If the image on the monitor is static, then it can be represented as the sum of voltages m – periodic i-th sequences of signals, for Fourier series coefficients:

$$s(t, \mathbf{v}_0) = \mathbf{1}(T_{a0} - t) \sum_{k=-\infty}^{\infty} \sum_{i=1}^{m} \frac{U_i}{\pi k} \sin\left(\frac{\pi k \tau_i}{T_f} \right) \cos\left[\frac{2\pi k}{T_f} \left[t - t_{di} \right] \right] \quad (6)$$

$\mathbf{1}(x)$ – single function, T_{a0} – true duration of the unchanged image on the monitor screen. As illustrated in Fig. 1 example $m = 3$

The zero harmonic is further excluded from consideration, since it does not propagate in space. Then (6) can be expressed in terms of amplitude quadratures:

$$a_{kc} = \sum_{i=1}^{m} \frac{2U_i}{\pi k} \sin\left(\frac{\pi k \tau_i}{T_f}\right) \cos\left(\frac{2\pi k t_{di}}{T_f}\right)$$

$$a_{ks} = \sum_{i=1}^{m} \frac{2U_i}{\pi k} \sin\left(\frac{\pi k \tau_i}{T_f}\right) \sin\left(\frac{2\pi k t_{di}}{T_f}\right)$$

in the following form:

$$s(t,\mathbf{v}) = \mathbf{1}(T_{a0} - t) \sum_{k=1}^{\infty} \left[a_{kc} \cos\left(\frac{2\pi k t}{T_f}\right) + a_{ks} \sin\left(\frac{2\pi k t}{T_f}\right) \right] \quad (7)$$

Since the periodic structure of the signal does not change over a period $t = [0, T_{a0}]$, each $a_{kc}(t)$ – and i $a_{ks}(t)$ – and amplitude quadrature with cyclic information leakage frequencies $2k\pi / T_f$ is constant: $a_{kc}(t) = \text{const1}$; $a_{ks}(t) = \text{const2}$, and change only when $t > T_{a0}$. Fig. 2 shows, for example, changes in time of the $a_{205c}(t)$-th, $a_{10023s}(t)$-th, and $a_{200045c}(t)$-th quadrature of amplitudes for three realisations of random time intervals T_a of changing static images on the monitor screen – $T_{a10}, T_{a20}, T_{a30}$.

As you can see, each k – and the quadrature amplitude is a sequence of rectangular pulses of unknown amplitude $a \in [0, a_{\max}]$ at intervals $T_{a10}, T_{a20}, T_{a30}, \ldots$, and substitutions of (7) in (5) for (2):

$$B(t,\tau) = \frac{N_0}{2} \delta(t - \tau)$$

where $\delta(\tau)$ – Delta function, N_0 – one-way spectral power density of white Gaussian noise (WGN), gives a solution to integral equation (4), for $t \in [0, T_a]$:

$$V(t,\mathbf{v}) = \frac{2}{N_0} \sum_{k=1}^{\infty} \left[a_{kc} \cos\left(\frac{2\pi k t}{T_f}\right) + a_{ks} \sin\left(\frac{2\pi k t}{T_f}\right) \right] \quad (8)$$

So, whatever the structure of the periodic (with a follow-up period T_f) signal $s(t,\mathbf{v})$, its energy is invested in quadrature amplitudes a_{kc} and a_{ks}, $k = 1, 2, .., \infty$:

$$\begin{aligned} E(\mathbf{v}) = \int_0^{T_{a0}} s^2(t,\mathbf{v})\,\mathrm{d}t = \frac{T_f}{2\pi} \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \frac{1}{k^2 - l^2} \times \\ \times \left(\left(a_{kc}a_{lc}k - a_{ks}a_{ls}l\right) \sin\left(\frac{2\pi k T_{a0}}{T_f}\right) \cos\left(\frac{2\pi l T_{a0}}{T_f}\right) + \right. \\ + \left(a_{ks}a_{ls}k - a_{kc}a_{lc}l\right) \cos\left(\frac{2\pi k T_{a0}}{T_f}\right) \sin\left(\frac{2\pi l T_{a0}}{T_f}\right) - \\ - \left(a_{kc}a_{ls}k + a_{ks}a_{lc}l\right) \sin\left(\frac{2\pi k T_{a0}}{T_f}\right) \sin\left(\frac{2\pi l T_{a0}}{T_f}\right) - \\ - \left(a_{ks}a_{lc}k + a_{kc}a_{ls}l\right) \cos\left(\frac{2\pi k T_{a0}}{T_f}\right) \cos\left(\frac{2\pi l T_{a0}}{T_f}\right) + \\ \left. + \left(a_{ks}a_{lc}k - a_{kc}a_{ls}l\right) \right) \end{aligned} \quad (9)$$

Substituting (8) in (4) allows getting the LRF:

$$\begin{aligned} L(T_a,\mathbf{v}) = \frac{2}{N_0} \sum_{k=1}^{\infty} \int_0^{T_a} x(t) \left[a_{kc} \cos\left(\frac{2\pi k t}{T_f}\right) + \right. \\ \left. + a_{ks} \sin\left(\frac{2\pi k t}{T_f}\right) \right] \mathrm{d}t - \frac{E(\mathbf{v})}{N_0}, \end{aligned} \quad (10)$$

in which $E(\mathbf{v})$ – is set (9).

The convergence of the double row in (9) depends on the specific "image" on the monitor screen. But in the vast majority of cases, in (10) we can limit ourselves to the total number of information harmonics with frequencies k / T_f, k – harmonic number, $k = 1, 2, \ldots, K$, which reaches $K \approx 3.3 \cdot 10^7$, since for $T_f = 1/60$ Hz the upper frequency limit of IEMR&I – does not exceed 2 GHz, for most monitors on LCS of the computational facilities.

Such quantitative restriction of unknown parameters allows us to solve the problem of finding MLA for LRF, in which the vector of unknown signal parameters $\mathbf{v}$ is $2K$ – measurable:

$$\begin{aligned} L(T_a, a_{1c}, a_{1s}, a_{2c}, a_{2s}, .., a_{Kc}, a_{Ks}) = \frac{2}{N_0} \sum_{k=1}^{K} \int_0^{T_a} x(t) \cdot \\ \cdot \left[a_{kc} \cos\left(\frac{2\pi k t}{T_f}\right) + a_{ks} \sin\left(\frac{2\pi k t}{T_f}\right) \right] \mathrm{d}t - \\ - \frac{1}{N_0} \sum_{k=1}^{K} \int_0^{T_a} \left[a_{kc} \cos\left(\frac{2\pi k t}{T_f}\right) + a_{ks} \sin\left(\frac{2\pi k t}{T_f}\right) \right]^2 \mathrm{d}t \end{aligned} \quad (11)$$

The essence of MLA is that unknown $2K$ – quadrature amplitudes a_{kc}, a_{ks}, $k = 1, 2, \ldots, K$, replace them with the most plausible estimates – $\hat{a}_{kc}, \hat{a}_{ks}$, which are arguments of alternately solved equations:

$$\begin{aligned} \frac{\partial L(T_a, a_{1c}, a_{1s}, a_{2c}, a_{2s}, .., a_{Kc}, a_{Ks})}{\partial a_{kc}} = 0 \\ \frac{\partial L(T_a, a_{1c}, a_{1s}, a_{2c}, a_{2s}, .., a_{Kc}, a_{Ks})}{\partial a_{ks}} = 0 \end{aligned} \quad (12)$$

The arguments of solutions (12) for quadrature amplitudes will have the form:

$$\begin{aligned} \hat{a}_{kc} \approx \frac{2}{T_a} \int_0^{T_a} x(t) \cos\left(\frac{2\pi k t}{T_f}\right) \mathrm{d}t \\ \hat{a}_{ks} \approx \frac{2}{T_a} \int_0^{T_a} x(t) \sin\left(\frac{2\pi k t}{T_f}\right) \mathrm{d}t \end{aligned} \quad (13)$$

which are more accurate than $T_a >> T_f$ and then the signal-to-noise ratio at the output of the linear part of the STMI receiver is greater, and substituting (13) in (11) allows us to obtain LRF for MLA, in which the uncertainty with respect to quadrature amplitudes is overcome:

$$\begin{aligned} L(T_a) = \frac{4}{N_0 T_a} \sum_{k=1}^{K} \left[\left(\int_0^{T_a} x(t) \cos\left(\frac{2\pi k t}{T_f}\right) \mathrm{d}t \right)^2 + \right. \\ \left. + \left(\int_0^{T_a} x(t) \sin\left(\frac{2\pi k t}{T_f}\right) \mathrm{d}t \right)^2 \right] \end{aligned} \quad (14)$$

A compatible algorithm for detecting IEMR&I and estimating the time interval at which they do not change consists in comparing the absolute maximum of (14) with the detection threshold h, in which a decision is made to detect IEMR&I from the monitor screen when

$$\sup_{T_{a\min} \le T_a \le T_{a\max}} L(T_a) > h \quad (15)$$

and the decision that there is no IEMR&I signal – when

$$\sup_{T_{a\min} \le T_a \le T_{a\max}} L(T_a) \le h$$

A diagram of the asymptotically Bayesian compatible IEMR&I signal detection algorithm and image duration estimation is shown in Fig. 3.

It consists of K – energy storage channels k harmonics in quadratures, K – adders in processing channels from $T_{a\min}$ to $T_{a\max}$, maximum selection device (MSD) and a comparison scheme with the detection threshold h. Since estimate of the frame duration on the monitor screen $\hat{T}_{a0}$ is known, it is possible, by using (13), to obtain estimates of k spectral quadrature amplitudes (see Fig. 4).

Diagrams of Fig. 3 and Fig. 4 fully describe the possibilities of modern STMI with asymptotically optimal algorithms for joint detection of signals and evaluation of their informative parameters, except for the procedure for estimating the period of following the frame scan of the monitor on LCS, which precedes the detection of the IEMR&I signal from computational facilities. To reproduce a black-and-white image intercepted by the STMI from the screen of the monitor of computational facilities, it is enough to generate a signal (7) using amplitude estimates (13) and feed it to the monitor's video card (after mixing all RGB signals).

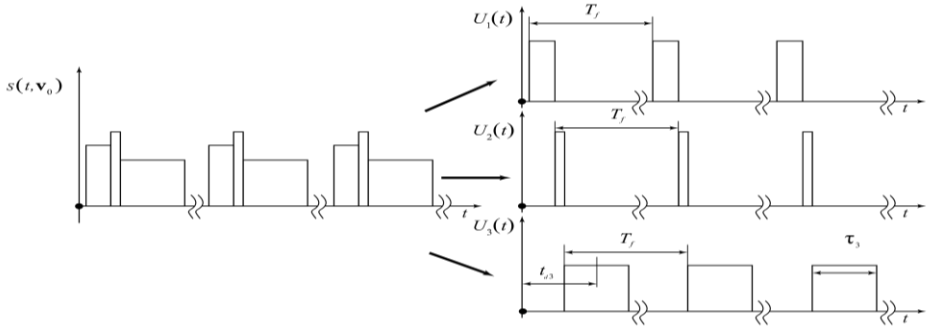


Fig. 1. Decomposition into elementary sequences of rectangular pulses of the information signal of the monitor on LCS

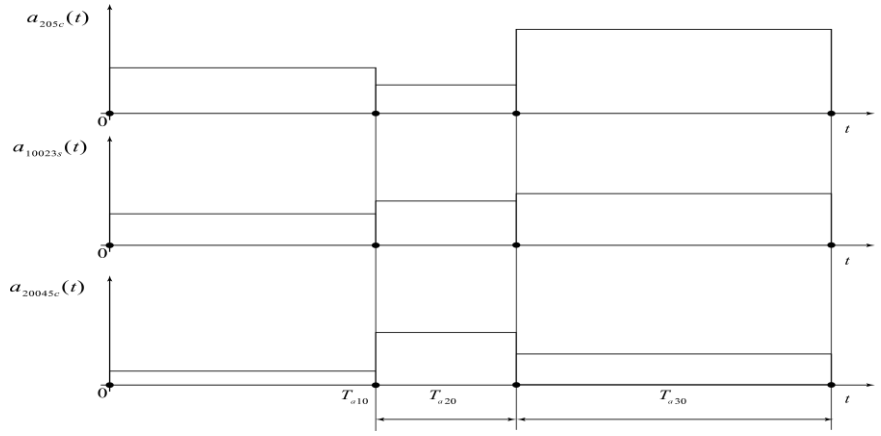


Fig. 2. Examples of changes in quadrature amplitudes over time

3. Further ways to solve the problem

The synthesis of STMI, which better intercepts static images from the monitor screen om LCS, is necessary to assess the potential capabilities of the enemy by "reading" information from the monitor screen. The better the IEMR&I signal detection quality indicators, for example, according to the Neumann-Pearson criterion: the greater the probability of correct signal detection

$$D = 1 - F_0(h, T_a/1) \tag{16}$$

in case of a fixed false alarm

$$1 - F_0(h, T_a/0) \tag{17}$$

the worse the information is protected. In (16) and (17) $F_0(x, T_a/0)$, $F_0(x, T_a/1)$ – distribution of absolute maxima (DAM) of the process (10), when the IEMR&I signal is present in the implementation $x(t)$ and when it is not, respectively.

However, each specific solution to the problem of analysing STMI quality indicators will depend on a specific "picture" on the monitor screen and a specific vector of parameters $\mathbf{v}_0 = (a_{1c0}, a_{1s0}, a_{2c0}, a_{2s0}, .., a_{Kc0}, a_{Ks0})$, and all further tasks of analysing the quality of detection of IEMR&I signals of STMI are reduced to the search for DAM $F_0(x, T_a/0)$, $F_0(x, T_a/1)$ random process (10), found only for most discontinuous signals.

Since the process at the output of a linear system is limited by frequency K/T_f, its spectral power density at the output of a linear system when there is no IEMR&I signal at the input of the STMI, taking into account (1) and, will be equal to:

$$S_o(T_a, T_f, \omega) = \frac{N_0}{2} \times \begin{cases} \left| \dfrac{\sin\left(\dfrac{\omega T_f}{2}\left(2\,\mathrm{int}\left[\dfrac{T_a}{2T_f}\right]+1\right)\right)}{\omega \sin\left(\dfrac{\omega T_f}{2}\right)} \right|, & 0 \le \omega \le 2\pi K / T_f \\ 0, & \omega > 2\pi K / T_f \end{cases} \tag{18}$$

where $\mathrm{int}[x]$ – integer part of a number x. The corresponding (18) correlation function presented by the Taylor series for small τ, and is equal to:

$$B(T_a, T_f, \tau) = \frac{1}{2\pi} \int_0^{2\pi K/T_f} S_o(T_a, T_f, \omega)\left(1 - \frac{\omega^2 \tau^2}{2}\right) d\omega = \sigma^2(T_a, T_f) - \frac{\sigma^2(T_a, T_f)\,\omega_2(T_a, T_f)\,\tau^2}{2} \tag{19}$$

where the variance of the non-stationary process is equal to:

$$\sigma^2\left(T_a,T_f\right)=\frac{1}{2\pi}\int_0^{2\pi K/T_f} S_o\left(T_a,T_f,\omega\right)\mathrm{d}\omega$$

and the second spectral moment of the process is equal to:

$$\omega_2\left(T_a,T_f\right)=\frac{\int_0^{2\pi K/T_f} S_o\left(T_a,T_f,\omega\right)\omega^2\mathrm{d}\omega}{\int_0^{2\pi K/T_f} S_o\left(T_a,T_f,\omega\right)\mathrm{d}\omega}$$

The correlation function of the process (19) corresponds to a differentiated in the mean square non-stationary process, for which the search for a DAM $F_0\left(x,T_a/0\right)$ and $F_0\left(x,T_a/1\right)$ has been a scientific problem until recently. Solution of the analysis problem for such processes, in cases where further spectral moments can be considered $\omega_4\left(T_a,T_f\right)$, $\omega_6\left(T_a,T_f\right)$, .., $\omega_{2n}\left(T_a,T_f\right)$ – infinite, is presented in the monograph [5]. However, the practice of applying the developed mathematical apparatus proves that even in the case of finite higher spectral moments, this does not significantly affect the results of the analysis. It is enough to fulfil condition $\omega_4\left(T_a,T_f\right)>>\omega_2^2\left(T_a,T_f\right)$ for everything to work satisfactorily [1–3].

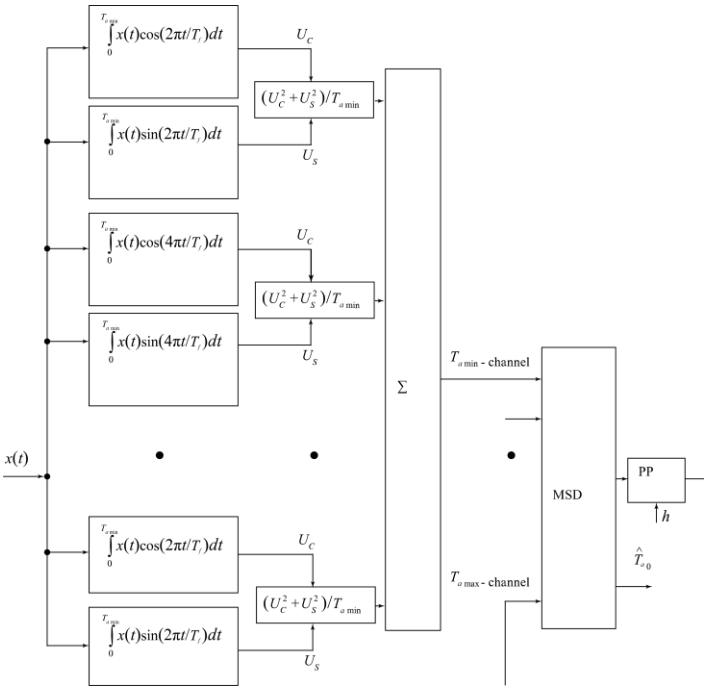


Fig. 3. Asymptotically Bayesian compatible algorithm for detecting IEMR&I and estimating the duration of a static image on a monitor screen

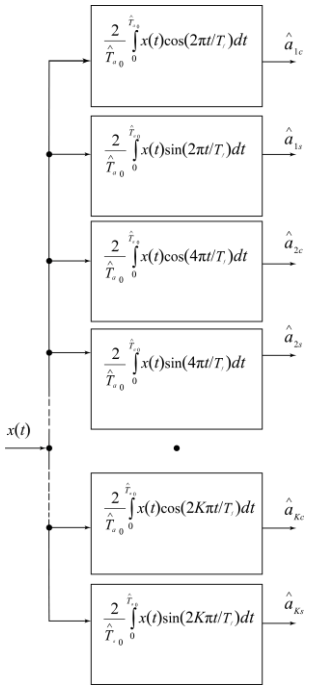


Fig. 4. Block diagram of the IEMR&I quadrature amplitude estimation device

4. Conclusions

1. Synthesis of the maximum likelihood algorithm in the form of asymptotically Bayesian detection of side electromagnetic radiation and interference signals is possible only when the number of unknown quadrature amplitudes of the leakage signal a_{kc}, a_{ks} is finite – $2K$.
2. The finite number of unknown quadrature amplitudes during the analysis of algorithms for detecting signals of adverse electromagnetic radiation from the monitor screen does not allow using the well-known apparatus of one-component Markov processes, suitable for breaking signals.
3. Further solution of the problem of analysing algorithms for detecting signals of side electromagnetic radiation and interference in a specialised enemy reconnaissance equipment requires finding distributions of absolute maxima of non-stationary processes differentiated in the mean square.

References

[1] Adler R. J.: Random field and geometry. Springer, New York 2007.
[2] Adler R. J.: Random field and their geometry. Springer, New York 2003.
[3] Adler R. J.: The geometry of random fields. J.Wiley & Sons, New York 1981.
[4] Adler R. J.: The geometry of random fields. Society for Industrial and Applied Mathematics, Philadelphia 2010.
[5] Kuhn M. G.: Security Limits of Compromising Emanations. Cryptographic Hardware and Embedded Systems. Journal 3659, 2005, 265–279.
[6] Leadbetter M. R.: Extreme and Related Properties of Random Sequences and Processes. Springer, New York 1983.
[7] Nakonechnyi T. A., Yevgrafov D. V.: Interception of an information leak signal from the monitor screen. International scientific and technical conference "Radio Engineering fields, signals, apparatuses and systems", Kyiv 2018, 45–52.
[8] Tanaka H., Takizawa O., Yamamura A.: A Trial of the interception of Display image using Emanation of Electromagnetic Wave. J. Inst. Image Electron. Eng. Jpn 34(1), 2005, 147–155.
[9] Yevgrafov D. V.: Physical foundations of information protection in radio-electronic equipment: textbook. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv 2014.
[10] Yevgrafov D. V.: Signal detection by means of electronic intelligence. monograph. Publishing house "Condor", Kyiv 2020.

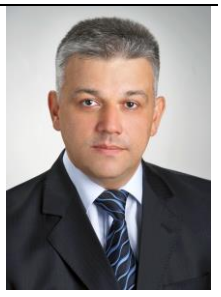
Ph.D. Dmytro V. Yevgrafov
e-mail: ramgraf@bigmir.net

The Ph.D. (Engineering), and the Senior Researcher. He is the author of more than 50 scientific academic papers, including 2 monographs, 45 articles in scientific specialized publications, 4 training manuals, materials and abstracts of reports at academic conferences. The author's scientific interests are in statistical radio engineering, radio electronic intelligence and physical protection of information.

http://orcid.org/0000-0001-9651-1558

D.Sc. Eng. Yurii Ye. Yaremchuk
e-mail: yurevyar@vntu.edu.ua

The Doctor of Science in Engineering and the Professor. He is the author of more than 300 publications, including 2 monographs, 140 articles in scientific specialized publications, 20 textbooks and training manuals, 20 utility model patents and 25 certificates of copyright registration for work. The author's scientific interests are in cryptographic and steganographic protection of information, technical protection of information and security of information systems.



http://orcid.org/0000-0002-6303-7703

http://doi.org/10.35784/iapgos.5301 received: 14.09.2023 | revised: 30.10.2023 | accepted: 27.11.2023 | available online: 20.12.2023

AC POWER REGULATION TECHNIQUES FOR RENEWABLE ENERGY SOURCES

Mariusz Ostrowski
Wroclaw University of Science and Technology, Faculty of Electronics, Photonics and Microsystems, Wroclaw, Poland

Abstract. *This article explores different AC power regulation techniques that can be employed to optimize the output of renewable energy sources, such as solar and wind power systems. The article provides an overview of the challenges associated with regulating AC power output from renewable sources and examines various techniques that can be used to improve the performance of power regulation systems. These techniques include voltage control, phase control, reactive power compensation, and power factor correction. The article also discusses the benefits and limitations of each technique, as well as their potential applications in renewable energy systems. Overall, this article provides valuable insights for engineers and researchers working to optimize power auto consumption in renewable energy systems.*

Keywords: ac power regulators, pulse width modulation converters, renewable energy systems

TECHNIKI REGULACJI MOCY ODBIORNIKÓW AC DLA ODNAWIALNYCH ŹRÓDEŁ ENERGII

Streszczenie. *W tym artykule omówiono różne techniki regulacji mocy prądu zmiennego, które można wykorzystać do optymalizacji produkcji energii ze źródeł odnawialnych, takich jak systemy energii słonecznej i wiatrowej. Artykuł zawiera przegląd wyzwań związanych z regulacją mocy wyjściowej prądu przemiennego ze źródeł odnawialnych i analizuje różne techniki, które można wykorzystać do poprawy wydajności systemów odnawialnych źródeł energii. Techniki te obejmują sterowanie napięciem, sterowanie fazą, kompensację mocy biernej i korekcję współczynnika mocy. W artykule omówiono również zalety i ograniczenia poszczególnych technik, a także ich potencjalne zastosowania w systemach energii odnawialnej. Artykuł zawiera cenne informacje dla inżynierów i badaczy pracujących nad optymalizacją auto konsumpcji energii elektrycznej w systemach energii odnawialnej.*

Słowa kluczowe: regulatory mocy prądu przemiennego, przetwornice modulacji szerokości impulsu, systemy energii odnawialnej

Introduction

Renewable energy sources, such as solar and wind power plants, have become increasingly popular due to their numerous benefits over traditional fossil fuels. However, one of the main challenges in harnessing these energy sources is their inherent variability and intermittency. This means that their output power levels can fluctuate unpredictably over time, making it difficult to maintain a stable and consistent electricity supply to the grid.

To address this issue, various AC power regulation techniques have been developed to effectively manage fluctuations in renewable energy sources and ensure a stable power supply. These techniques involve the use of advanced power electronics and control systems that can efficiently convert, condition, and regulate the AC power generated from renewable energy sources.

This article provides an overview of some of the most common AC power regulation techniques used in self-consumption devices found in photovoltaic systems. The operating principles, advantages, and limitations of each technique will be discussed, along with their potential applications in various renewable energy systems. By the end of this article, readers will have a better understanding of how AC power regulation techniques can impact the quality of energy in self-consumption devices utilized within renewable energy systems.

1. The essence of power/voltage regulation of AC receivers powered from renewable energy sources

Voltage stability is an important parameter of the state of the power grid. According to the standards determined individually for each country, its value must be within the specified standards. For example, in Poland, according to the PN-EN 50160 standard, the correct voltage value should be within the range of 207-253V RMS (230V nominal).

In figure 1, a voltage graph for one of the three phases is presented. The measurement was conducted on April 11, 2023, using an Agilent 34461A multimeter, with measurements taken at 1-second intervals. The measurements were taken in a small town where several photovoltaic installations are installed. The graph shows numerous voltage fluctuations which can be associated with the activation of various devices. From the chart, it can also be observed that the voltage value is higher than the nominal voltage. This is due to the injection of energy into the grid by photovoltaic installations. During peak hours, the voltage value exceeds the permissible limits (253 V). This may be caused by improper settings of some inverters in residential photovoltaic installations. Sudden fluctuations in the grid voltage during hours of electricity overproduction can also be associated with the cascading shutdown of inverters due to the detection of excessively high voltage on any phase. From the perspective of energy suppliers, this is a significant issue, especially during the summer period.

Therefore, it appears necessary to use devices that increase self-consumption of electrical energy [2]. Typically, these are various types of domestic hot water tanks equipped with a high-power heater with electronic control system to regulate temperature. The heater is activated during the production of electrical energy by the photovoltaic installation. More expensive solutions are enhanced with power control systems that adapt it to the current production of the photovoltaic installation.

The second figure shows a voltage chart on each of the phases on May 21, 2023, recorded by the SoFar SF4ES005 inverter. From the chart, one can read The excessive voltage level and its fluctuations are independent on each phase. Three-phase inverters shut down energy production on all phases if voltage value standards are exceeded on even one of them. This results in abrupt shutdowns of photovoltaic installations, leading to sudden changes in grid voltage across all phases. This causes losses in energy production for prosumers and is another reason to implement advanced self-consumption systems with sophisticated control systems.

Another important parameter is the amount of additional harmonics introduced into the power grid. This is a very important parameter due to the need to reduce voltage distortions described in the electrical standards applicable in various countries (in Poland, PN-EN 61000). The most problematic harmonics are those with low frequencies due to issues related to designing LC filters or EMI filters [4]. The cutoff frequency of low-pass reactive (LC) filters is described by the equation [6]:

$$f_{gr} = \frac{1}{\pi\sqrt{LC}} \tag{1}$$

For a frequency of 50 Hz, both the required capacitance and inductance must have large values. Furthermore, to improve the amplitude response of the filter (for passive filters, the roll-off rate is typically 20dB/decade), it is necessary to combine them cascadingly. The higher the values of the first 10 harmonics, the more complex the problem becomes when designing a filter. Capacitors with high capacitance cause a phase shift between current and voltage, which in turn increases the reactive power component.

Each electricity receiver changes the parameters of the power grid to a greater or lesser extent. A particular problem for power grids are small photovoltaic installations in a prosumer system, and potentially new challenges in the form of high-power receivers where power regulation is necessary [1]. Therefore it is so important to choose the right control method for a specific solution.

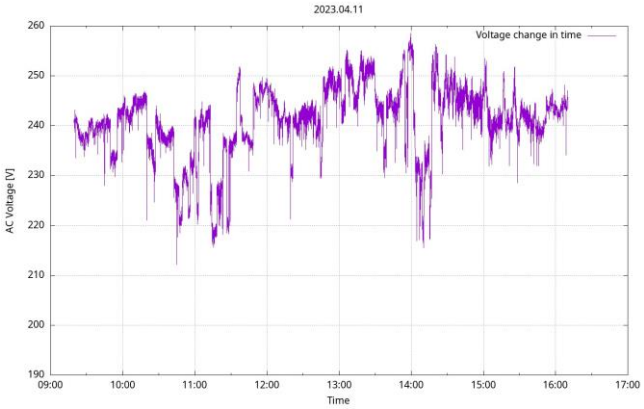


Fig. 1. Voltage waveform measured on one of the three phases made on April 11, 2023 using the Agilent 34461A multimeter

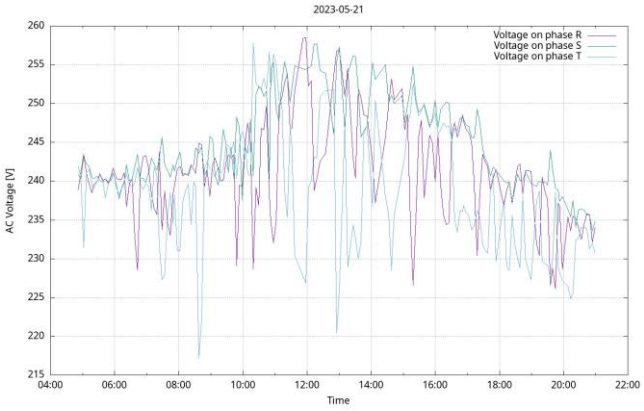


Fig. 2. Voltage waveform on each of tree phase reported by Sofar SF4ES005 inverteter on April 11, 2023

2. Power regulation techniques for AC loads

The power generated by renewable energy systems is highly variable over time. This is especially noticeable in photovoltaic systems when on a sunny day the output voltage is high enough that the inverter shuts down to meet the requirements of the grid. This causes a reduction in energy production and, at the same time, voltage spikes in the transmission grids. To prevent this, various types of systems that increase self-consumption of energy are used. They usually act as thermal energy storages in which electrical energy is converted by heaters into thermal energy. These heaters are used mainly in boilers to heat up domestic hot water. The simplest method of increasing self-consumption is a system in which the heater is simply switched on when the energy is sent to the grid. This may cause an unnecessary load on the transmission lines due to the incorrect adjustment of the heater power to the current production of electricity by photovoltaic panels. Therefore, in more advanced systems, the power of the energy receiver is regulated so that the total energy consumed by the house is equal to 0 if it posiible. Usually, power regulation is accomplished by altering the output voltage. This can be achieved by converting alternating current to direct current, and then regulating the voltage using a PWM controller or by converting it back to AC voltage using an inverter [5]. There are three main techniques for controlled power regulation of AC receivers [3].

2.1. Phase Angle Control

Phase Angle Control is a widely used method for regulating the power delivered to AC loads. This method involves adjusting the firing angle of a thyristor, triac or other solid-state switch in the AC circuit to control the amount of power delivered to the load. By adjusting the phase angle, the effective voltage and current delivered to the load can be controlled.

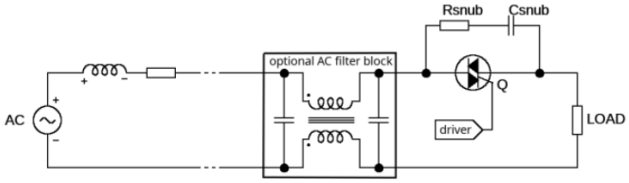


Fig. 3. Circuit layout of typical AC power regulator for Phase Angle Control or Burst Fire control

One of the primary advantages of Phase Angle Control is its simplicity and low cost. A typical circuit layout of the regulator is shown in Fig. 3. Compared to other methods such as PWM control, Phase Angle Control requires fewer components and can be implemented using relatively simple control circuits. Phase Angle Control method has good power resolution and fast response to set-point changes. Examples of the output voltage waveforms are shown in Fig. 4.

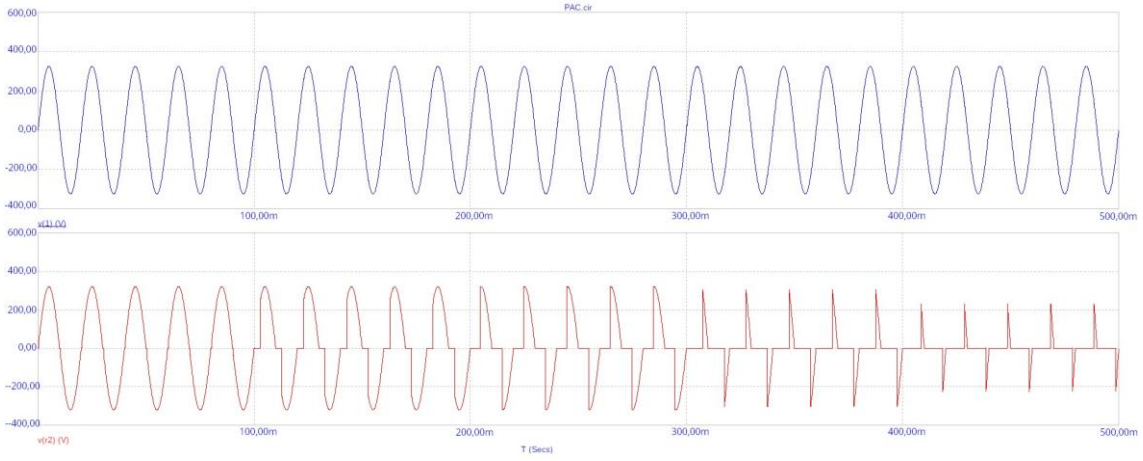


Fig. 4. Input voltage and load voltage waveforms for 100%, 75%, 50%, 25% load for Phase Angle Control method

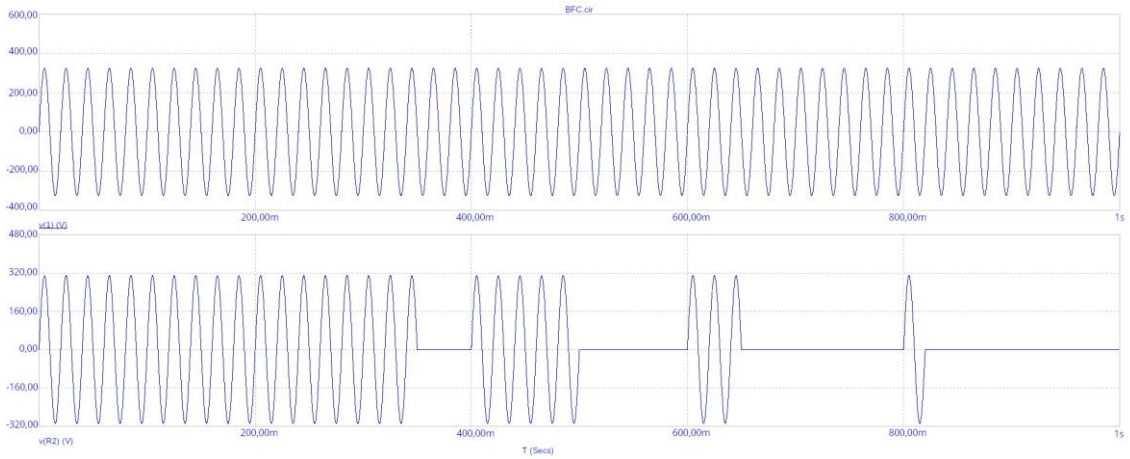


Fig. 5. Input voltage and load voltage waveforms for 100%, 75%, 50%, 25% load for Burst Fire Control method

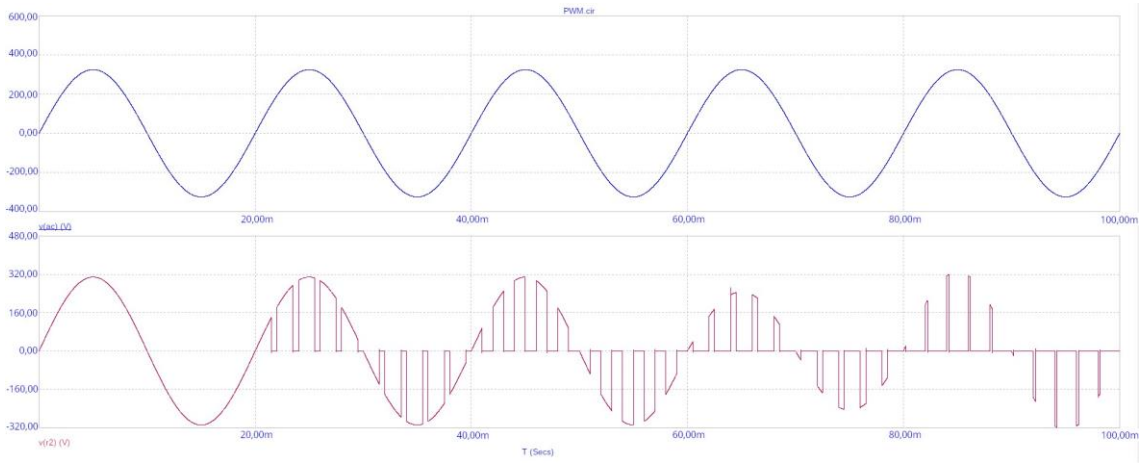


Fig. 7. Input voltage and load voltage waveforms for 100%, 75%, 50%, 25% and 10% duty cycle for PWM method

It can be seen that for a duty cycle below 50% the waveforms are heavily distorted. This can cause problems with the power supply of devices that are passive inductive or capacitive. Morever, there are some limitations to Phase Angle Control that should be considered. One major issue is the distortion of the AC waveform, which can lead to increased harmonic content and reduced power factor. To address these issues, several variations of Phase Angle Control have been developed, such as modified phase control and burst firing. These techniques involve adjusting the duration of the thyristor firing pulses to minimize distortion and improve efficiency.

2.2. Burst Fire Control

Burst Fire Control is a method of power regulation that is commonly used in AC power control applications. The circuit shown in Fig. 3 can be used for the regulation of output power. In Burst Fire Control method, the AC waveform is divided into a series of pulses, with each pulse being triggered by a separate thyristor, triac or other switching device. The duration of each pulse, or "burst", is controlled to regulate the power delivered to the load.

Fig. 5 shows the voltage waveform at the output of the Burst Fire Control for duty cycles of 100%, 75%, 50%, 25% and 10%.

One advantage of Burst Fire Control is its simplicity and cost-effectiveness. Compared to other AC power control methods, such as Phase-Angle Control, Burst Fire Control requires fewer components and less sophisticated control circuitry. This makes it a popular choice for applications where cost and simplicity are crucial factors. However, Burst Fire Control also comes with some disadvantages. This method can generate significant harmonic distortion in the input waveform, especially, which can cause issues with other equipment in the same electrical system. Moreover, due to its nature, it offers very low resolution. Additionally, this method can only be used for resistive loads.

2.3. Pulse Width Modulation

The basic principle of PWM (Pulse Width Modulation) regulators relies on cyclically turning on and off electronic switches, such as transistors or thyristors, to generate rectangular pulses. The ratio of the on-time of the switch to the total period determines the duty cycle of the PWM signal. A higher duty cycle corresponds to a longer pulse duration relative to the period, resulting in a higher effective voltage or current. The schematic diagram of an example PWM regulator is shown in Fig. 6.

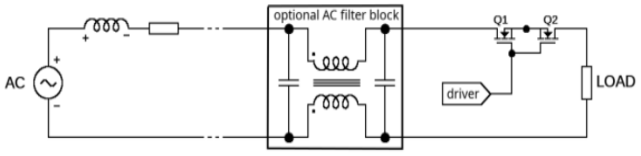


Fig. 6. Circuit layout of typical AC power regulator for PWM control

In the simplest version, the regulator consists of two MOSFET transisotrs switched on simultaneously. Compared to the Phase Angle Control regulator, it requires the use of a more complicated gate driver system. Fig. 7 shows the voltage waveform at the output of the pwm controller operating at 1 kHz for duty cycles of 100%, 75%, 50%, 25% and 10%.

The PWM control regulator operates at a high frequency (above 30kHz), thanks to which disturbances in the power grid are easier to filter out. Morever, the PWM controller does not cause flickering as in the case of the Burst Fire Control controller. This type of controllers can work with resistive loads, but more importantly, after applying appropriate LCL filters, also with inductive or capacitive loads.

3. Simulation results

The simulations were conducted using Microcap 12 software. In the simulations, the impact of power regulation methods on distortions appearing in the power grid was compared. For this purpose, a resistive load with a rated power of 2 kW was connected to the control system. The research was carried out for 6 regulation cases: 25%, 50%, 75%, and 100% for voltages ranging from 217V to 253V RMS (in accordance with PN-EN 50160).

Charts 8-10 depict Harmonic Distortion Analysis plots for regulators at 50% load power regulation and voltages ranging from 217V to 253V RMS. The analysis presents the first 10 harmonics appearing on the utility side of the power grid.

In figure 8, the result of the Phase Angle Control regulator's operation is presented. The third harmonic of the signal – 150 Hz, has an amplitude of 5.24V, which corresponds to 1.7% of the grid voltage value. The successive harmonics had the following values: 5th – 1.6 V, 7th – 1.6 V, 9th – 1.2 V (THD 1.74%).

In figure 9, the result of the Burst Fire Control regulator's operation is presented. The third harmonic of the signal – 150 Hz, has an amplitude of 2.1 μV, 5th – 0.12 μV, 7th – 1.2 μV, 9th – 1.4 μV (THD 3μ%). These values are significantly smaller compared to the pulse angle method, which greatly simplifies, and in some cases, completely eliminates the need for reactive filters.

In figure 10, the result of the Pulse Width Modulation regulator's operation is presented. The third harmonic of the signal – 150 Hz, has an amplitude of 22.2 mV, 5th – 35.5 mV, 7th – 12.4 mV, 9th – 15.7 mV (THD 2.25%).

The above analysis indicates that the amplitude of the dominant harmonics does not change with variations in the electrical network voltage in the range of 217 V to 253 V RMS. From the conducted analysis, it can be inferred that among the discussed methods, the Burst Fire Control method is the best. This directly results from the fact that in this method, the load is always switched on for the full period. Analyzing only the 3rd harmonic, this method generates distortions that are 1000 times smaller than the PWM method. The Phase Angle Control method exhibits the worst parameters. The results of the harmonic distortion analysis for the remaining controller settings are presented in table 1.

The analysis indicates that the lowest harmonic distortions are generated by the Burst Fire Control controller. Regardless of the power regulation setting, the amplitude of the first 10 harmonics does not exceed 3 μV. In the case of the PWM controller, only in the case of 100% power setting does it exhibit slightly lower distortion levels than the others. This is due to its more complex control system compared to the other controllers and the ability to continuously activate the load. For the other settings, the amplitude of the first 10 harmonics does not exceed 40 mV.

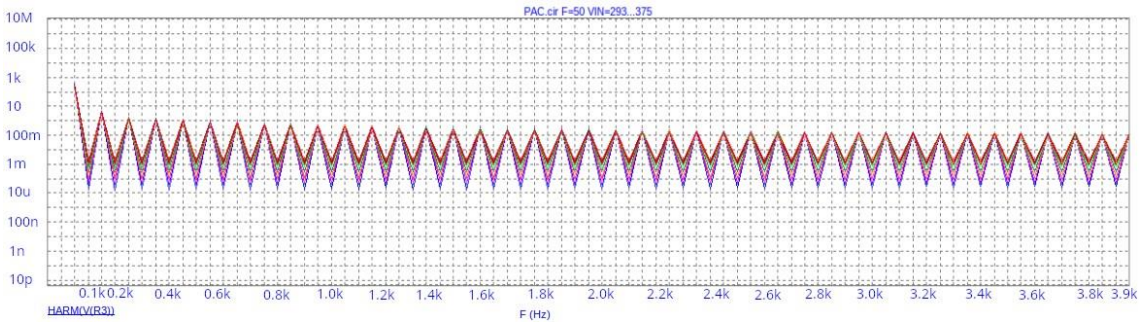


Fig. 8. Harmonic distortion analysis for the Phase Angle Control regulator at 50% load power regulation and voltages ranging from 217V to 253V RMS

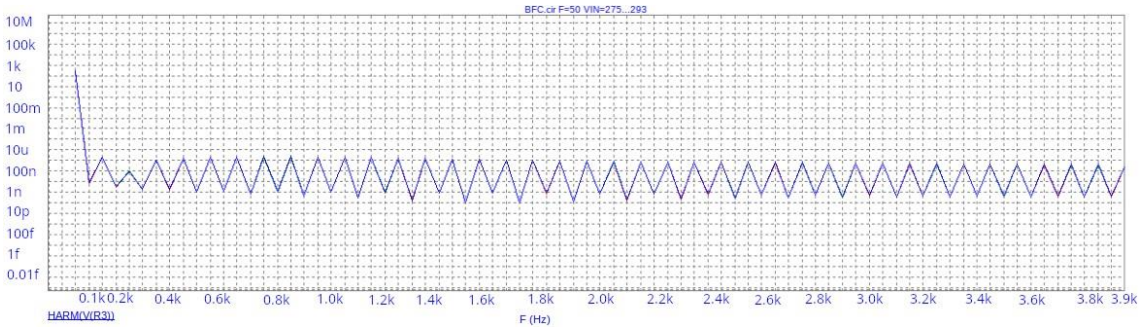


Fig. 9. Harmonic distortion analysis for the Burst Fire Control regulator at 50% load power regulation and voltages ranging from 217V to 253V RMS

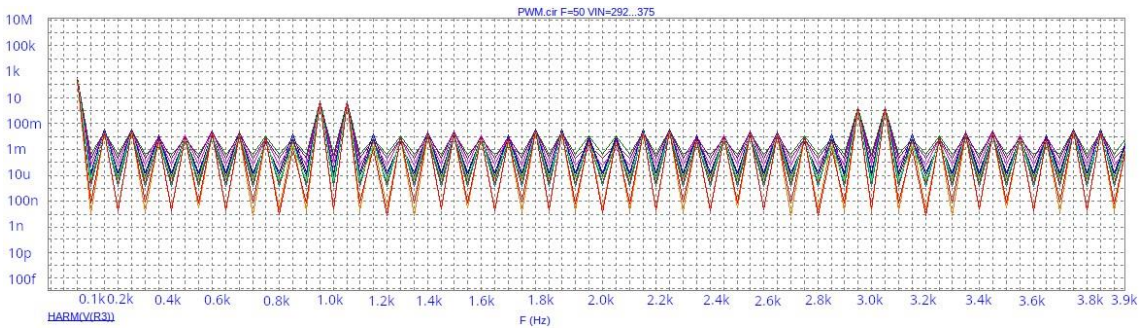


Fig. 10. Harmonic distortion analysis for the PWM control regulator at 50% load power regulation and voltages ranging from 217V to 253V RMS

Table 1. Results of the harmonic distortion analysis for each of the discussed power regulation methods at load percentages of 25%, 50%, 75%, and 100%

Harmonic	3th [V]	5th [V]	7th [V]	9th [V]	max
25%					
Phase Angle Control	4.59	2.2	1.2	1.04	3th – 4.6 V
Burst Fire Control	2.1µ	0.12µ	1.2µ	1.4µ	3th – 2.1 µV
Pulse Width Modulation	19.1m	17.1m	21.8m	18.8m	19th – 3.1 V
50%					
Phase Angle Control	5.24	1.6	1.6	1.2	3th – 5.2 V
Burst Fire Control	2.1µ	0.12µ	1.2µ	1.4µ	3th – 2.1 µV
Pulse Width Modulation	22.2m	35.3m	12.4m	15.7m	19th – 3.5 V
75%					
Phase Angle Control	2.3	2	1.2	0.64	3th – 2.3 V
Burst Fire Control	2.1µ	0.12µ	1.2µ	1.4µ	3th – 2.1 µV
Pulse Width Modulation	31.6m	28.4m	21.3m	31.6m	19th – 3.4 V
100%					
Phase Angle Control	93m	89m	89m	86m	3th – 93 mV
Burst Fire Control	2.1µ	0.12µ	1.2µ	1.4µ	3th – 2.1 µV
Pulse Width Modulation	2.4µ	0.4µ	0.1µ	0.1µ	3th – 2.4 µV

Table 2. Results of the harmonic distortion analysis for each of the discussed power regulation methods at load percentages of 25%, 50%, 75%, and 100% after applying the EMI filter

Harmonic	3th [V]	5th [V]	7th [V]	9th [V]	max
25%					
Phase Angle Control	191m	32.1m	6.44m	2.26m	3th – 191 mV
Burst Fire Control	1µ	0.01µ	0.18µ	0.26µ	3th – 1 µV
Pulse Width Modulation	551µ	56.5µ	38.3µ	22.6µ	19th – 42 mV
50%					
Phase Angle Control	273m	26.6m	8.5m	3m	3th – 273 mV
Burst Fire Control	1µ	0.01µ	0.18µ	0.26µ	3th – 1 µV
Pulse Width Modulation	421µ	68µ	28µ	21.8µ	19th – 56 mV
75%					
Phase Angle Control	174.1m	28m	6m	2.1m	3th – 174 mV
Burst Fire Control	1µ	0.01µ	0.18µ	0.26µ	3th – 1 µV
Pulse Width Modulation	395u	64µ	30µ	10.7µ	19th – 37 mV
100%					
Phase Angle Control	2.8m	0.68m	0.27m	0.13m	3th – 2.8 mV
Burst Fire Control	1.1µ	0.01µ	0.18µ	0.26µ	3th – 1 µV
Pulse Width Modulation	1.4µ	0.01µ	0.01µ	0.001µ	3th – 1.4 µV

On figure 11, the waveform of one cycle of the electrical power network voltage is presented for the pulse angle control method with the load set at 50% for voltages ranging from 217 V to 253 V RMS. Distortions with low-frequency characteristics can be observed. If the number of devices connected to the same phase is increased, the problem will exacerbate. In the case of the PWM method, as shown in figure 12, distortions appear evenly throughout the entire period, and their frequency depends solely on the switching frequency of the controller. This is an advantage of this control method. Moreover, with a greater number of devices operating at different frequencies, there should be no increase in the amplitude of the first harmonics of the electrical power network voltage.

Figure 13 shows the voltage waveform on one phase for the burst fire method. In the graph, you can observe the appearance of distortions with a much lower frequency than 50 Hz. These distortions result from the low-frequency switching of the switches in the controller. These disturbances can cause noticeable flickering of the lighting.

Table 2 presents the results of the analysis after applying an optional low-pass filter as shown in figures 3 and 6. The calculated inductance value for the PWM controller operating at a frequency of 1 kHz was 2.2 mH, and the capacitance was 50 µF. By increasing the operating frequency of the controller circuit to frequencies above 20 kHz (typically, controllers operate at frequencies greater than 30 kHz), the capacitance value can be reduced to 1 µF. For the Pulse Angle controller, the required capacitance value is 1 mF with an inductance of 10 mH.

The application of the filter reduced the amplitude of the first 10 harmonics by a factor of 100 to 1000 in the case of a PWM controller. In the case of the Pulse Angle method, the amplitude decreased 20–100 times. However, in the case of the Burst Fire Control method, the use of a low-pass EMI filter does not change the amplitude of harmonics appearing on the power grid side.

For low-frequency pulsations caused by the operation of the Burst Fire Control controller, the use of high-pass filters with a cut-off frequency of around 1 Hz is required, which may not be practical from a design and application perspective. In the case of the other two methods, the use of a filter significantly contributed to the reduction of harmonics.

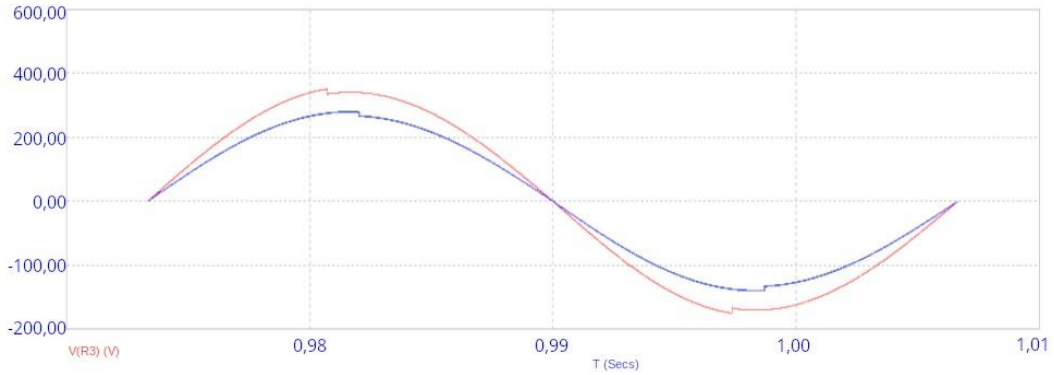


Fig. 11. Waveform of one cycle of the electrical power network voltage for the pulse angle control method with the receiver's power set to 50%

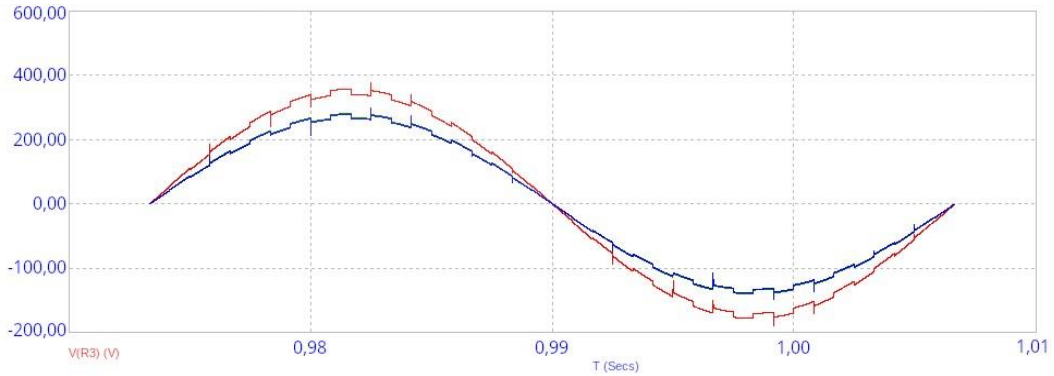


Fig. 12. Waveform of one cycle of the electrical power network voltage for the PWM method with the receiver's power set to 50%

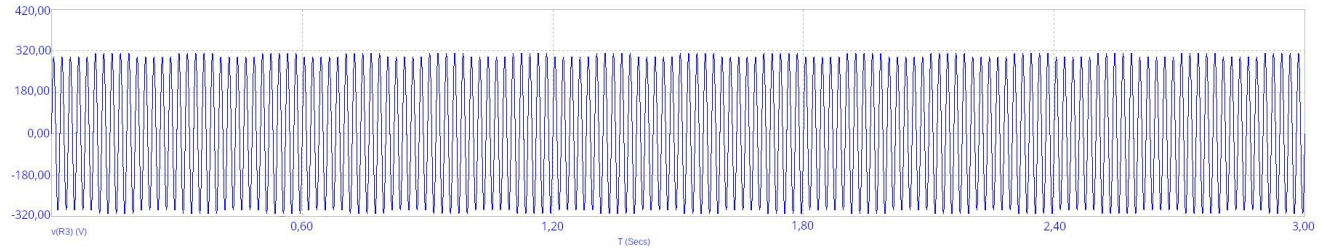


Fig. 13. Voltage ripple of the electrical power network for the Burst Fire Control method with a 50% load

4. Summary

The choice of the appropriate power regulation method for a receiver is crucial from the perspective of the power grid. The right method should not introduce additional distortions, and a proper control system will ensure that the voltage does not exceed permissible values. By using power regulation systems, it is possible to significantly increase energy self-consumption without drawing additional energy from the power grid.

For high-power systems, the PWM method appears to have the best parameters. After applying the appropriate filters, with significantly lower capacitance and inductance compared to other methods, the disturbances propagated to the power grid are the smallest among all the discussed techniques. Furthermore, by increasing the switching frequency of the switches, the size of the EMI filter can be reduced. This method also does not induce low-frequency pulsations, surpassing the Burst Fire Control method.

In summary, the PWM method has the best parameters for applications in power control of receivers operating in systems that increase energy self-consumption in photovoltaic systems.

References

[1] Calpa M. et al.: Effects of Large-scale PV Self-consumption on the Aggregated Consumption. Procedia Computer Science 83, 2016, 816–823.
[2] Masa-Bote D. et al.: Improving photovoltaics grid integration through short time forecasting and self-consumption. Applied Energy 125, 2014, 103–113.
[3] McDonald D. W.: AC power control in the Core Flow Test Loop. No. ORNL/TM--7139. Oak Ridge National Lab., 1980.
[4] Ozenbaugh R. L. et al.: EMI filter design. CRC press, 2000.
[5] Tzou Y. Y., Shih-Liang J.: Full control of a PWM DC-AC converter for AC voltage regulation. IEEE Transactions on aerospace and electronic systems 34(4), 1998, 1218–1226.
[6] Williams A. B., Taylor F. J.: Electronic filter design handbook. McGraw-Hill Education, 2006.

Ph.D. Eng. Mariusz Ostrowski
e-mail: mariusz.ostrowski@pwr.edu.pl

Since 2013 lecturer of the Department of Electronic and Photonic Metrology at the Wrocław University of Science and Technology. In 2019 he obtained a doctorate in engineering and technical sciences in the field of electronics
His research interests focus on renewable energy sources, mainly on systems for optimizing the generation and consumption of energy from photovoltaic panels.

http://orcid.org/0000-0001-6797-9880

http://doi.org/10.35784/iapgos.5377 received: 26.07.2023 | revised: 01.09.2023 | accepted: 24.11.2023 | available online: 20.12.2023

AUTOMATIC ADJUSTMENT OF REACTIVE POWER BY FACTS DEVICES UNDER CONDITIONS OF VOLTAGE INSTABILITY IN THE ELECTRIC NETWORK

Mykhailo Burbelo[1], Oleksii Babenko[1], Yurii Loboda[1], Denys Lebed[1], Oleg K. Kolesnytskyj[1], Saule J. Rakhmetullina[2], Murat Mussabekov[3]

[1]Vinnytsia National Technical University, Vinnytsia, Ukraine, [2]D.Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan, [3]Institute "Transport Engineering" of JSC "Academy of Logistics and Transport", Almaty, Kazakhstan

Abstract. *This article describes the problem of automatic regulation of reactive power using electronic devices FACTS (Flexible AC Transmission Systems): static synchronous compensator (STATCOM) and unified power flow controller (UPFC). With the help of a complex writing form, the following are determined: voltages at the installation nodes of the FACTS device and loads, currents of loads, power sources and electronic compensators in case of voltage instability at the load node of the electrical network. Voltages and currents are determined using the node-voltage method. The task of STATCOM is partial or full compensation of reactive power. During the reduction of the voltage at the load node, the reactive power generated by the power source decreases. The STATCOM should partially or fully compensate for the reactive power imbalance as quickly as possible. However, at the same time, it is not possible to fully compensate for the voltage reduction. A series-parallel or parallel-series UPFC can be used to solve this problem. As a result of using the UPFC, it is possible to automatically raise the voltage level to acceptable values with the help of the UPFC series compensator. The analysis shows that the parallel-serial UPFC is characterized by the stability of operation. In the case of using a series-parallel UPFC, there are restrictions on the ability to adjust the imaginary voltage component of the series compensator, since the angle of the voltage vector changes, which causes a failure in the operation of the regulator of the parallel compensator UPFC.*

Keywords: reactive power, static synchronous compensator, unified power flow regulator, voltage instability

AUTOMATYCZNA REGULACJA MOCY BIERNEJ PRZEZ URZĄDZENIA FACTS W WARUNKACH NIESTABILNOŚCI NAPIĘCIA W SIECI ELEKTRYCZNEJ

Streszczenie. *W artykule opisano problematykę automatycznej regulacji mocy biernej za pomocą urządzeń elektronicznych FACTS (Flexible AC Transmission Systems): statycznego kompensatora synchronicznego (STATCOM) oraz regulatora przepływu mocy (UPFC). Za pomocą złożonego formularza rejestracyjnego określane są: napięcia w węzłach instalacji urządzenia FACTS i obciążenia, prądy obciążenia, źródeł zasilania i kompensatorów elektronicznych w przypadku niestabilności napięcia w węźle obciążenia sieci elektrycznej. Napięcia i prądy są wyznaczane metodą napięć węzłowych. Zadaniem STATCOM jest częściowa lub pełna kompensacja mocy biernej. Podczas spadku napięcia w węźle obciążenia, moc bierna generowana przez źródło zasilania maleje. STATCOM powinien częściowo lub w pełni skompensować nierównowagę mocy biernej tak szybko, jak to możliwe. Jednocześnie jednak nie jest możliwe pełne skompensowanie spadku napięcia. W celu rozwiązania tego problemu można zastosować szeregowo-równoległy lub równoległo-szeregowy układ UPFC. W wyniku zastosowania UPFC możliwe jest automatyczne podniesienie poziomu napięcia do akceptowalnych wartości za pomocą kompensatora szeregowego UPFC. Analiza pokazuje, że równoległo-szeregowy UPFC charakteryzuje się stabilnością działania. W przypadku zastosowania szeregowo-równoległego UPFC istnieją ograniczenia w zakresie możliwości regulacji składowej urojonej napięcia kompensatora szeregowego, ponieważ zmienia się kąt wektora napięcia, co powoduje awarię działania regulatora kompensatora równoległego UPFC.*

Słowa kluczowe: moc bierna, statyczny kompensator synchroniczny, ujednolicony regulator przepływu mocy, niestabilność napięcia

Introduction

In modern conditions, there is a tendency to increase reactive power with a variable character, which is a problem for the effective use of alternating current power lines [7]. Conventional (not equipped with FACTS devices) power grids have the problem of low utilization of transmission line capacity, which reduces their efficiency due to limited possibilities of regulating reactive power flows [12].

FACTS devices are used to solve these problems: a static synchronous compensator (STATCOM) and a unified power flow controller (UPFC) [5]. The main advantage of STATCOM and UPFC is the automatic regulation of reactive power flows in electrical networks. The UPFC, in addition, provides an increase in the voltage level at the load. The UPFC adjustable parameters are line impedance, phase angle, and voltage. UPFC allows performing such an important function as ensuring voltage stability. The use of UPFC regulators ensures the improvement of the static and dynamic characteristics of power transmission lines [4, 5, 11].

UPFC is a combination of a Static Series Synchronous Compensator (SSSC) and Static Parallel Compensator (STATCOM) that share a DC capacitor (Fig. 1) [1, 11,15]. The series compensator improves the dynamic stability of the network load node in transient mode, the parallel compensator provides dynamic regulation of reactive power flows [8,13,14].

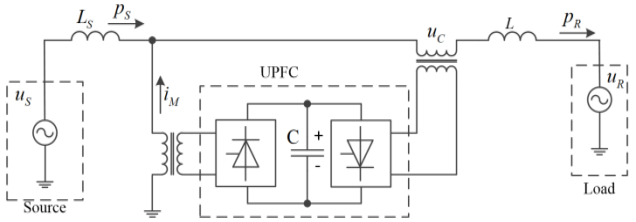


Fig. 1. Scheme UPFC

1. STATCOM simulation in static mode

The most widely used FACTS device for reactive power compensation is the STATCOM, which is designed for smooth regulation of reactive power within wide limits (±100%), has high speed in the modes of consumption and generation of reactive power, has relatively small dimensions, can be used in almost any electrical networks. The use of STATCOM allows not only to regulate reactive power, but also to increase network bandwidth, and optimize power flows and voltage levels. Its use in "weak" networks to reduce voltage deviations and fluctuations is particularly effective.

Fig. 2 shows the scheme of replacing a network node with a STATCOM.

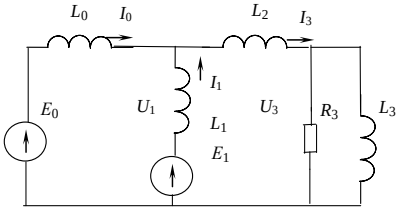


Fig. 2. Scheme of replacing a network node with STATCOM

The following values of network node parameters in relative units are stated as follows: $X_0 = 0.15$; $X_1 = 0.015$; $X_2 = 0.1$; $X_3 = 2$; $R_3 = 1$. Conductivity of branches

$$Y_0 = \frac{1}{jX_0}; Y_1 = \frac{1}{jX_1}; Y_2 = \frac{1}{jX_2}; Y_3 = \frac{1}{R_3} + \frac{1}{jX_3} \quad (1)$$

Voltages U_1 and U_3 are calculated according to the formula:

$$\begin{bmatrix} U_1 \\ U_3 \end{bmatrix} = \begin{bmatrix} Y_0 + Y_1 + Y_2 & -Y_2 \\ -Y_2 & Y_2 + Y_3 \end{bmatrix} \begin{bmatrix} Y_0 \cdot E_0 + Y_1 \cdot E_1 \\ 0 \end{bmatrix} \quad (2)$$

The E_1 STATCOM parameter is selected in the adjustment process in such a way that the active power of the compensator is equal to zero, and the reactive power provides the optimal mode in the electrical network. For $E_0 = 1.05$, $E_1 = 1.0 - j0.13$ values of voltages: $U_1 = 0.995 - j0.129$ and $U_3 = 0.928 - j0.212$. The currents, accordingly, of the power source, the compensator and the load are equal:

$$I_0 = Y_0 \cdot (E_0 - U_1) = 0.863 - j0.364$$
$$I_1 = Y_1 \cdot (E_1 - U_1) = -0.041 - j0.311$$
$$I_3 = Y_3 \cdot U_3 = 0.822 - j0.675$$

The powers, accordingly, of the power source, the compensator and the load will be:

$$S_0 = E_0 \cdot \overline{I_0} = 0.906 + j0.383$$
$$S_1 = E_1 \cdot \overline{I_1} = 0.000 + j0.316$$
$$S_3 = E_3 \cdot \overline{I_3} = 0.906 + j0.453$$

The reactive power of the load and transmission lines is compensated by the power source and STATCOM.

Fig. 3 shows the amplitude-phase characteristics of currents I_0 and I_1 in the event of a decrease in the voltage at the power supply node in the range from $1.05 \cdot U_н$ to $0.9 \cdot U_н$. The arrow indicates the direction of change in the current components in case of a decrease in voltage.

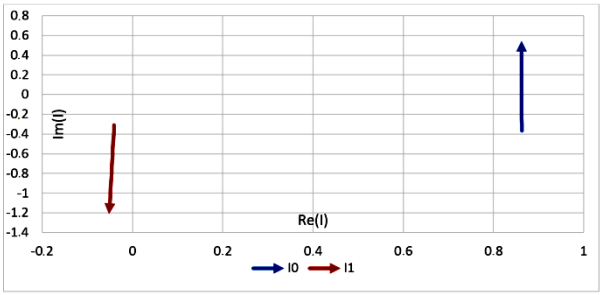


Fig. 3. Amplitude-phase characteristics of currents I_0 and I_1 in the event of a voltage drop in the power node with STATCOM

A decrease in the voltage of the power source leads to a decrease in the reactive component of the current I_0, and then to a change in its direction. The power supply reduces the generation of reactive power and then begins to consume it. In this case, the decrease in the reactive power of the power supply is compensated by the STATCOM.

To reduce the sharp decrease in reactive power consumption by the power source, it is necessary to reduce the effective EMF component of the compensator E_1 to the level of 0.95. However, as can be seen from Fig. 4, the active component of the current I_1 increases slightly, which characterizes the additional consumption of the active power of the STATCOM.

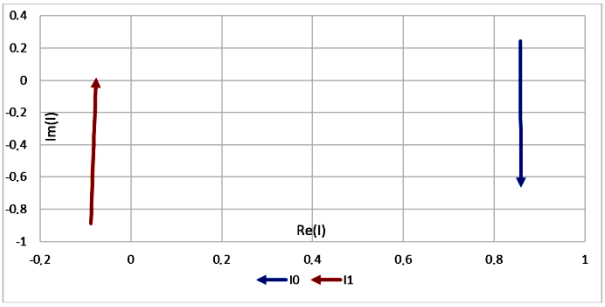


Fig. 4. Amplitude-phase characteristics of currents I_0 and I_1 in case of reduction of the real component E1 STATCOM

At the same time, the voltage level on the load remains insufficient. In the case of reducing the voltage of the power supply to the value $E_0 = 1.0$ the load voltage is $U_3 = 0.924 - j0.211$.

2. Simulation of parallel-serial UPFC in static mode

UPFC can be used to increase the load voltage. Fig. 5 shows the scheme of replacing a network node with a parallel-serial UPFC [6, 11].

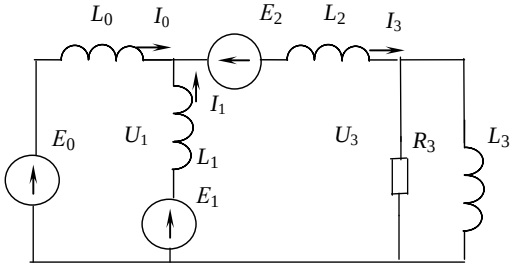


Fig. 5. Scheme of replacing a network node with a parallel-serial UPFC

Voltages U_1 and U_3 calculated according to the formula:

$$\begin{bmatrix} U_1 \\ U_3 \end{bmatrix} = \begin{bmatrix} Y_0 + Y_1 + Y_2 & -Y_2 \\ -Y_2 & Y_2 + Y_3 \end{bmatrix} \begin{bmatrix} Y_0 \cdot E_0 + Y_1 \cdot E_1 + Y_2 \cdot E_2 \\ -Y_2 \cdot E_2 \end{bmatrix} \quad (3)$$

For $E_0 = 1.05$, $E_1 = 1.0 - j0.155$ and $E_2 = -0.1 - j0.15$ values of voltages:

$$U_1 = 0.996 - j0.154,\ U_3 = 1.034 - j0.103$$

currents:

$$I_0 = Y_0 \cdot (E_0 - U_1) = 1.029 - j0.359$$
$$I_1 = Y_1 \cdot (E_1 - U_1) = -0.046 - j0.260$$
$$I_3 = Y_3 \cdot U_3 = 0.983 - j0.620$$

powers:

$$S_0 = E_0 \cdot \overline{I_0} = 1.080 + j0.377$$
$$S_1 = E_1 \cdot \overline{I_1} = -0.006 + j0.267$$
$$S_2 = -E_2 \cdot \overline{I_3} = 0.005 + j0.209$$
$$S_3 = E_3 \cdot \overline{I_3} = 1.080 + j0.540$$

Reactive power is compensated by the power supply, parallel and series UPFC compensators [9, 10].

In the case of reducing the voltage of the power supply to the value $E_0 = 1.0$ voltage on the load $U_3 = 1.030 - j0.102$ practically does not change; the power of the series compensator does not change either $S_2 = 0.005 + j0.209$. In this case, the decrease in reactive power of the power source is compensated by a parallel compensator.

Fig. 6 shows the amplitude-phase characteristics of currents I_0 and I_1 in the event of a voltage drop in the power supply node in the range of $1.05 \ldots 0.9 \cdot U_n$.

A decrease in the voltage of the power source, as in the scheme with a STATCOM, leads to a decrease in the reactive component of the current I_0, and then to a change in the sign of the reactive component of the current I_0 (Fig. 7). The power source begins to consume reactive power, and the UPFC parallel compensator increases its generation, ensuring the balancing of reactive power at the load node.

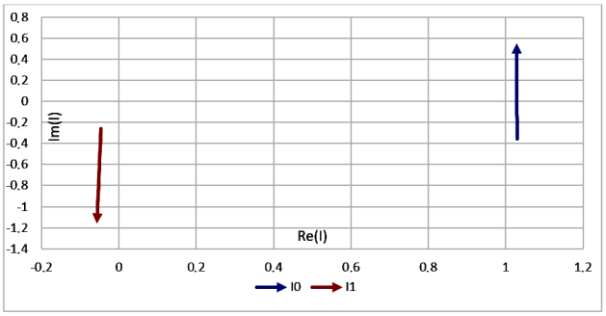


Fig. 6. Amplitude-phase characteristics of currents I_0 and I_1 in the event of a voltage drop in the power supply node with UPFC

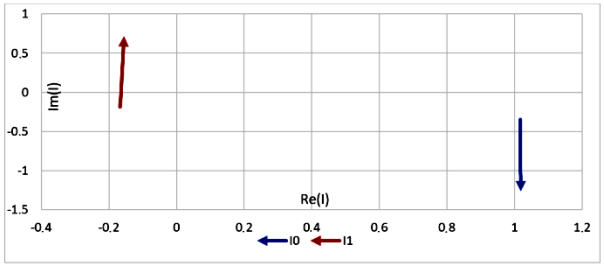


Fig. 7. Amplitude-phase characteristics of currents I_0 and I_1 in the event of a decrease in the real component E_1

To reduce the sharp decrease in reactive power consumption by the power source, it is necessary to reduce the effective EMF component of the parallel compensator E_1 to the level of 0.85. However, as can be seen from Fig. 7, the active component of current I_1 increases, which characterizes the additional consumption of active power by the compensator. The additional consumption of active power by the compensator can be compensated by simultaneously adjusting both EMF components without changing the phase angle, i.e. by adjusting the EMF module.

3. Modelling of series-parallel UPFC in static mode

Fig. 8 shows a substitution scheme with series-parallel UPFC. Voltages U_1 and U_3 are calculated according to the formula:

$$\begin{bmatrix} U_1 \\ U_3 \end{bmatrix} = \begin{bmatrix} Y_0 + Y_1 + Y_2 & -Y_2 \\ -Y_2 & Y_2 + Y_3 \end{bmatrix} \begin{bmatrix} Y_0 (E_0 - E_2) + Y_1 E_1 \\ 0 \end{bmatrix} \quad (4)$$

For $E_1 = 1.0 - j0.145$, $E_2 = -0.1$ values of voltages:

$$U_1 = 1.094 - j0.144 \text{ and } U_3 = 1.02 - j0.234$$

currents:

$$I_0 = Y_0 \cdot (E_0 - E_2 - U_1) = 0.961 - j0.371$$

$$I_1 = Y_1 \cdot (E_1 - U_1) = -0.058 - j0.374$$

$$I_3 = Y_3 \cdot U_3 = 0.903 - j0.744$$

powers:

$$S_0 = E_0 \cdot \overline{I_0} = 0.961 + j0.371$$

$$S_1 = E_1 \cdot \overline{I_1} = -0.010 + j0.419$$

$$S_2 = -E_2 \cdot \overline{I_1} = -0.006 + j0.037$$

$$S_3 = E_3 \cdot \overline{I_3} = 1.095 + j0.548$$

Reactive power is compensated by a power supply and a parallel compensator.

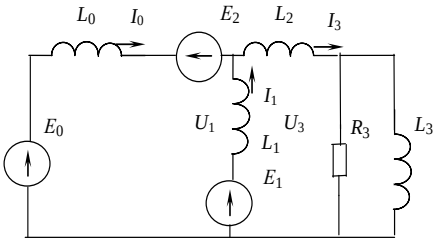


Fig. 8. Scheme of substitution of a network node with a series-parallel UPFC

In the case of reducing the voltage of the power supply to the value $E_0 = 1.0$ voltages $U_1 = 1.090 - j0.144$ and $U_3 = 1.016 - j0.234$ practically do not change.

However, in the case of a slight change in the imaginary voltage component of the series compensator $E_2 = -0.1 - j0.05$ the consumption of the active power of the parallel compensator increases sharply $I_1 = -0.362 - j0.671$, and in the case of changing the sign of the imaginary component $E_2 = -0.1 + j0.05$ the generation of active power increases sharply $I_1 = 0.239 - j0.678$.

Such instability limits the use of the series-parallel UPFC scheme.

4. UPFC modeling in dynamic mode

Fig. 9 shows the functional diagram of the control system [2], which is proposed to be used to control the UPFC.

A model with a three-phase non-linear load was used to simulate UPFC operation [3]. Characteristics of the elements of the experimental model: resistance of the power supply system 18.4+*j*17.1 mΩ, resistance on the direct current side of the non-linear load (three-phase rectifier) – $12 + j1.256\ \Omega$, resistance of a symmetrical three-phase load – $10 + j10\ \Omega$. An inductance connected in series with SSSC $L_{se} = 0.5$ mH, with STATCOM $L_{sh} = 0.4$ mH. Two capacitors connected in series: capacitance $C = 2500$ μF, voltage $u_{dc} = 325$ V, $u_{dcref} = 650$ V. Filter parameters that are enabled in parallel to each of the blocks $C_{se} = 47$ μF, $C_{sh} = 10$ μF. Power of transformers $T_{se} = 2.5$ kV·A, $T_{sh} = 2.5$ kV·A.

Fig. 10,a shows graphs of the rms value of the supply voltage, which varies with the coefficient $k = 1.1, 1.09 \ldots 0.9$. Source voltage 400 V (230 V – RMS voltage of one phase). Researches are carried out using three-phase series-parallel UPFC and parallel-series UPFC. After the research, it was found that both regulators behaved identically and the voltage change graphs coincided. The behavior of SSSC (Fig. 10,b) has a slight offset (230.7 V) relative to the actual voltage value of 231 V when the power supply voltage deviates. An increase in the voltage of the power supply practically does not affect the operation of the SSSC.

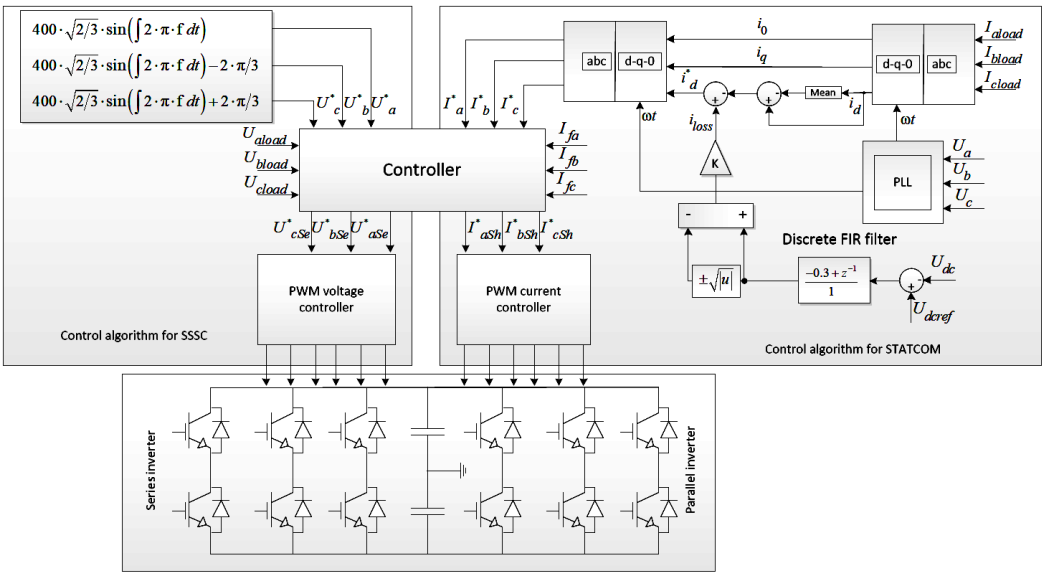


Fig. 9. Functional scheme of the UPFC control system

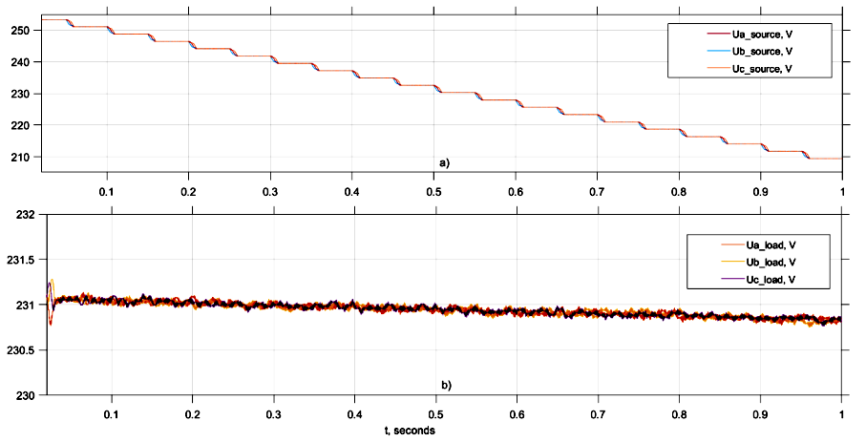


Fig. 10. The graph of the change in the rms value of the voltage: a) networks; b) loads

The graphs shown in Fig. 11a, b, c characterize, accordingly, the active P and reactive Q power of the power source, load and parallel compensator UPFC.

From these graphs, it is possible to determine: speed, power overshoot and the ability to compensate reactive power with UPFC.

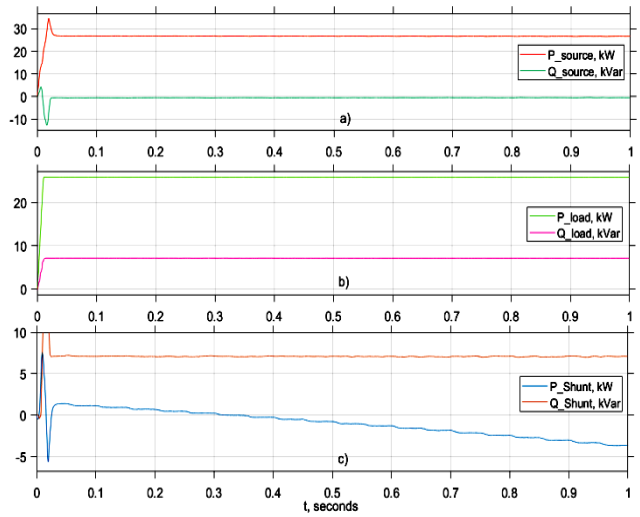


Fig. 11. The graph of the change in the rms value of the power: a) network; b) load; c) parallel compensator UPFC

The linearity of the power graphs depends on the power of the parallel transformer and the capacity of the DC storage capacitor. In case of an inappropriate setting, reactive power pulsation may occur.

The graph of the active power of the UPFC parallel compensator (Fig. 11c) shows the consumption of active power by the device at the coefficient $k = 1.1,\ 1.09 \ldots 1$ and its generation after reducing the input voltage coefficient to $k = 0.99,\ 0.98 \ldots 0.9$.

5. Conclusions

The behavior of STATCOM and UPFC has been demonstrated and the proposed control model has been verified under significant voltage disturbances at the load node of the electric network. STATCOM provides the fastest possible partial or full compensation of reactive power imbalance. However, at the same time, it is not possible to fully compensate for the reduction in voltage. A series-parallel or parallel-series UPFC can be used to solve this problem. As a result of using the UPFC, it is possible to automatically raise the voltage level to acceptable values with the help of the UPFC series compensator. The analysis shows that the parallel-serial UPFC is characterized by the stability of operation. In the case of using a series-parallel UPFC, there are restrictions on the ability to adjust the imaginary voltage component of the series compensator, since the angle of the voltage vector changes, which causes a failure in the operation of the regulator of the parallel compensator UPFC.

References

[1] Bian J., Lemak T. A., Nelson R. J., Ramey D. G.: Power Flow Controller Models for Power System Simulations. Power System Technology 19(9), 1995, 15–19.

[2] Burbelo M. Y., Lebed D. Yu.: Investigation of the Efficiency of the Capacitor Voltage Control Circuit of a Unified Electricity Quality Regulator to Reduce Voltage Fluctuations. Visnyk VPI 1, 2023, 21–28 [http://doi.org/10.31649/1997-9266-2023-166-1-21-28].

[3] Burbelo M. Y., Lebed D. Yu., Leshchenko O. R.: Optimization of charge / discharge time of active filter capacitors during voltage fluctuations. Herald of Khmelnytskyi national university 4, 2022, 58–63 [http://doi.org/10.31891/2307-5732-2022-311-4-58-63].

[4] Cañizares C. A., Uzunovic E., Reeve J.: Transient Stability and Power Flow Models of the Unified Power Flow Controller for Various Control Strategies. International Journal of Energy Technology and Policy 4(3-4), 2006, 349–378.

[5] Hingorani N. G., Gyugyi L.: Understanding FACTS. Concepts and Technology of Flexible AC Transmission Systems. IEEE Press book, 2000.

[6] Huang Z., Ni Y., Shen C. M., Wu F. F., Chen S., Zhang B.: Application of Unified Power Flow Controller in Interconnected Power Systems–Modeling, Interface, Control Strategy and Case Study. IEEE Trans. Power Systems 15(2), 2000, 817–824,

[7] Lee H.-J., Lee D.-S., Yoon Y.-D.: Unified Power Flow Controller Based on Autotransformer Structure. Electronics 8, 2019, 1542 [http://doi.org/10.3390/electronics8121542].

[8] Lezhniuk P., Komar V., Rubanenko O.: Information Support for the Task of Estimation the Quality of Functioning of the Electricity Distribution Power Grids with Renewable Energy Source. IEEE 7th International Conference on Energy Smart Systems – ESS 2020, 2020, 168–171.

[9] Lezhniuk P., Kravchuk S., Netrebskiy V., Komar V., Lesko V.: Forecasting Hourly Photovoltaic Generation on Day Ahead. IEEE 6th International Conference on Energy Smart Systems – ESS 2019, 2019, 184–187.

[10] Lezhniuk P., Kravchuk S., Buslavets O.: Selfoptimization Modes of Electric Grids with Renewable Energy Sources Using the Principle of Least Action. IEEE 6th International Conference on Energy Smart Systems – ESS 2019, 2019, 33–36.

[11] Mihalic R., Zunko P., Povh D.: Improvement of Transient Stability Using Unified Power Flow Controller. IEEE Trans. Power Delivery 11(1), 1996, 485–491.

[12] Padiyar K. R., Kulkarni A. M.: Control Design and Simulation of Unified Power Flow Controller. IEEE Trans. Power Delivery 13(4), 1998, 1348–1354.

[13] Papic I., Zunko P., Povh D.: Basic Control of Unified Power Flow Controller. IEEE Trans. Power Systems 12(4), 1997, 1734–1739.

[14] Schauder C. D., Gyugyi L., Lund M. R., Hamai D. M., Rietman T. R., Torgerson D. R., Edris A.: Operation of the Unified Power Flow Controller (UPFC) Under Practical Constraints. IEEE Trans. Power Delivery 13(2), 1998, 630–639.

[15] Wang S., Han L., Chen K.: Comprehensive coordinated control strategy of virtual synchronous generators under unbalanced power grid. J. Power Electron. 2019, 19, 1554–1565.

D.Sc. Eng. Mykhailo J. Burbelo
e-mail: burbelomj@gmail.com

Professor, Head of the Department of Electrical Systems of Power Consumption and Energy Management, Vinnytsia National Technical University. Research interests include: 1. Power supply (dynamic compensation of reactive power and optimization of power quality). 2. Dynamic measurements (measurement of parameters of electric circuits and modes of electric networks). He has published more than 200 scientific and educational and methodological works, including 6 monographs, 12 author's certificates and patents for inventions.

http://orcid.org/0000-0002-4510-2911

Ph.D. Oleksii V. Babenko
e-mail: oleksij_babenko@ukr.net

Candidate of Sciences (Engineering), associate professor of the Department of Electrical Engineering Systems of Power Consumption and Energy Management. He has published more than 30 scientific works. Field of scientific interests: improving the quality of electricity, current and voltage balancing installations of electrical network nodes and their measuring channels.

http://orcid.org/0000-0003-2773-6571

Ph.D. Yurii V. Loboda
e-mail: lobodaeseem@gmail.com

Senior lecturer of the Department of Electrical Systems of Power Consumption and Energy Management of the Vinnytsia National Technical University. Field of scientific interests: quality of electricity; improvement of the quality of electricity, management of devices for dynamic compensation of reactive power; increasing the reliability of distribution networks. He has published more than 20 scientific works. The main areas of scientific activity are improving the quality of electric power.

http://orcid.org/0000-0002-5397-4656

M.Sc. Denys Y. Lebed
e-mail: 4e15b.lebyd@gmail.com

Assistant of the Department of Electrical Systems of Power and Energy Management, Vinnytsia National Technical University. Field of scientific interests: problems of compensation of voltage fluctuations and flicker indicator; development of an active filter management system. Published 10 scientific works.

http://orcid.org/0000-0001-7482-9741

Ph.D. Oleg K. Kolesnytskyj
e-mail: kolesnytskiy@vntu.edu.ua

Ph.D. (Eng.), associate professor, Department of Computer Sciences, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia Field of scientific interests: information technologies, quality of electricity; improvement of the quality of electricity, management of devices for dynamic compensation of reactive power; increasing the reliability of distribution networks. He has published more than 100 scientific works.

http://orcid.org/0000-0003-0336-4910

Ph.D. Saule J. Rakhmetullina
e-mail: Rakhmetullinas@mail.ru

She is currently a Chairman of the Board – rector of the D. Serikbayev East Kazakhstan Technical University. She is a co-author over 40 papers in journals, book chapters, and conference proceedings and has 3 copyright certificates for software products. She is a Republican expert on reforming higher education. Her professional interests are mathematical and computer modeling of complex processes.

http://orcid.org/0000-0002-3142-0249

Ph.D. Murat Mussabekov
e-mail: murat-shar@mail.ru

M. Mussabekov works as an associate professor of the Department "Rolling Stock" of the Institute "Transport Engineering" of JSC "Academy of Logistics and Transport" Almaty, Kazakhstan. Advisor of Tengizchevroil Railway Department.
He is the author and co-author of textbooks, methodical manuals and more than 180 articles in journals and conference proceedings. Research area: Improvement of energy and environmental characteristics of automotive power plants.

http://orcid.org/0009-0004-1218-9696

http://doi.org/10.35784/iapgos.5375 *received: 26.07.2023 | revised: 25.09.2023 | accepted: 22.11.2023 | available online: 20.12.2023*

VENTILATION CONTROL OF THE NEW SAFE CONFINEMENT OF THE CHORNOBYL NUCLEAR POWER PLANT BASED ON NEURO-FUZZY NETWORKS

Petro Loboda[1], Ivan Starovit[1], Oleksii Shushura[2], Yevhen Havrylko[1], Maxim Saveliev[3], Natalia Sachaniuk-Kavets'ka[4], Oleksandr Neprytskyi[5], Dina Oralbekova[6], Dinara Mussayeva[7]

[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Department of Software Engineering in Energy, Kyiv, Ukraine, [2]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Department of Digital Technologies in Energy, Kyiv, Ukraine, [3]Institute for Safety Problems of Nuclear Power Plants National Academy of Sciences of Ukraine, Chornobyl, Ukraine, [4]Vinnytsia National Technical University, Vinnytsia, Ukraine, [5]Vinnytsia Mykhailo Kotsiubynskyi State Pedagogical University, Vinnytsia, Ukraine, [6]Satbayev University, Almaty, Kazakhstan, [7]Al Farabi Kazakh National University, Almaty, Kazakhstan

Abstract. *The accident at the Chornobyl Nuclear Power Plant (ChNPP) in Ukraine in 1986 became one of the largest technological disasters in human history. During the accident cleanup, a special protective structure called the Shelter Object was built to isolate the destroyed reactor from the environment. However, the planned operational lifespan of the Shelter Object was only 30 years. Therefore, with the assistance of the international community, a new protective structure called the New Safe Confinement (NSC) was constructed and put into operation in 2019. The NSC is a large and complex system that relies on a significant number of various tools and subsystems to function. Due to temperature fluctuations and the influence of wind, hydraulic processes occur within the NSC, which can lead to the release of radioactive aerosols into the environment. The personnel of the NSC prevents these leaks, including through ventilation management. Considering the long planned operational term of the NSC, the development and improvement of information technologies for its process automation is a relevant task. The purpose of this paper is to develop a method for managing the ventilation system of the NSC based on neuro-fuzzy networks. An investigation of the current state of ventilation control in the NSC has been conducted, and automation tools for the process have been proposed. Using an adaptive neuro-fuzzy inference system (ANFIS) and statistical data on the NSC's operation, neuro-fuzzy models have been formed, which allows to calculate the expenses of the ventilation system using the Takagi-Sugeno method. The verification of the proposed approaches on a test data sample demonstrated sufficiently high accuracy of the calculations, confirming the potential practical utility in decision-making regarding NSC's ventilation management. The results of this paper can be useful in the development of digital twins of the NSC for process management and personnel training.*

Keywords: New Safe Confinement, ventilation management, neuro-fuzzy network, information technology, fuzzy logic, digital twin

KONTROLA WENTYLACJI NOWEJ BEZPIECZNEJ POWŁOKI CZARNOBYLSKIEJ ELEKTROWNI JĄDROWEJ OPARTA NA ROZMYTYCH SIECIACH NEURONOWYCH

Streszczenie. *Awaria w Czarnobylskiej Elektrowni Jądrowej (ChNPP), która miała miejsce w Ukrainie w 1986 roku, stała się jedną z największych katastrof technologicznych w historii ludzkości. Podczas likwidacji awarii zbudowano specjalną strukturę ochronną – Obiekt "Ukrycie", mającą na celu izolację zniszczonego reaktora od otoczenia. Jednak planowany okres eksploatacji sarkofagu "Ukrycie" wynosił tylko 30 lat, dlatego przy wsparciu społeczności międzynarodowej zbudowano nową strukturę ochronną – "Nowa Bezpieczna Powłoka" (NSC), która została oddana do użytku w 2019 roku. NSC jest dużym i skomplikowanym systemem, którego funkcjonowanie zapewnia znaczna liczba różnych narzędzi i podsystemów. Ze względu na zmienne temperatury i wpływ wiatru, w NSC zachodzą procesy hydrauliczne, które mogą prowadzić do uwolnienia promieniotwórczych aerozoli do otoczenia. Personel NSC zapobiega tym wyciekom, między innymi poprzez zarządzanie wentylacją. W związku z długim planowanym okresem eksploatacji NSC, istotnym zadaniem jest rozwój i doskonalenie technologii informatycznych dla automatyzacji procesów. Celem pracy jest opracowanie metody zarządzania systemem wentylacji NSC opartej na rozmytych sieciach neuronowych. Przeprowadzono badania istniejącego stanu rozwiązywania problemów zarządzania wentylacją NSC oraz wybrano narzędzia do automatyzacji procesu. Za pomocą adaptacyjnego systemu wnioskowania neuro-rozmytego (ANFIS) i danych statystycznych dotyczących funkcjonowania NSC, stworzono modele neuro-rozmyte, które pozwalają na kalkulację kosztów systemu wentylacyjnego metodą Takagi-Sugeno. Weryfikacja zaproponowanych podejść na próbce kontrolnej danych wykazała wystarczająco wysoką dokładność obliczeń, co potwierdza możliwość ich praktycznego zastosowania w procesie podejmowania decyzji dotyczących zarządzania wentylacją NSC. Wyniki pracy mogą być również przydatne przy tworzeniu cyfrowe bliźniaków NSC w celu zarządzania procesami i szkolenia personelu.*

Słowa kluczowe: Nowa Bezpieczna Powłoka, zarządzanie wentylacją, rozmyta sieć neuronowa, technologia informacyjna, logika rozmyta, cyfrowy bliźniak

Introduction

As known, in 1986, an accident occurred at the Chernobyl Nuclear Power Plant (ChNPP), which became one of the largest man-made disasters in modern human history. To protect the population and the environment around the destroyed fourth power block of the ChNPP, a special structure was built, known as the Shelter Object (SO). However, the operational lifespan of the SO was only 30 years. Therefore, with the support of the international community, a new protective structure called the New Safe Confinement (NSC) was constructed and put into operation in Ukraine in 2019. One of the purpose of creating the NSC for the ChNPP is to protect the environment during the removal of radioactive materials from the destroyed fourth power block and the dismantling of its unstable structures [7].

The construction of the NSC provided isolation of the SO from the surrounding environment and created appropriate conditions for transforming it into an environmentally safe system. The NSC can be characterized as a large and complex system equipped with various equipment. The planned operational lifespan of the NSC is 100 years [4, 9]. To manage the processes of the NSC and monitor its condition, a specialized multi-level automated system has been developed and used.

However, the need to improve information visualization regarding the state of the NSC, ensure decision support, and automate the management of its processes necessitates the relevance of developing new information technologies. Taking into account the long operational lifespan of the NSC, approaches have been proposed for developing its IT architecture based on concepts such as "Continuous Development", "Continuous Integration", and "Continuous Deployment" [5].

Among the main tasks of managing the processes of the NSC is preventing releases of air from its main space with radioactive aerosols through existing construction leaks caused by temperature differentials and non-stationary hydraulic processes influenced by wind. Modeling and research have been conducted on the thermogasdynamic and radiation conditions of the NSC and SO, as well as their uncontrolled air exchange with the surrounding environment [8]. The model utilizing Computational Fluid Dynamics (CFD) models and artificial neural networks has been proposed to estimate the volumes of uncontrolled air releases with radioactive aerosols from the NSC into the environment under arbitrary wind speeds and directions [6, 7]. An internal pressure assessment model has been created for the NSC, along with approaches to manage hydraulic flows in order to minimize uncontrolled releases of air with radioactive aerosols when the input ventilation systems

are disconnected, thus reducing energy consumption [11]. The concept for a prediction and early detection system for the location and concentration of radioactive aerosols within the NSC has been developed [5], and the methodology has been presented for assessing the concentration of radioactive aerosols during operations related to the removal of fuel-containing materials from the SO [1]. It should be noted that the mentioned results are focused on individual processes of the NSC and do not involve the implementation of comprehensive state management.

One possible solution of this problem is the application of digital twin technology. In general terms, a digital twin is a virtual prototype of a real physical object, product, group of products, or process that collects and reuses digital information [2]. An architecture of the digital twin of the NSC at the ChNPP has been proposed, which includes a system of its models and algorithms [10]. A component of the digital twin development for the NSC is the modelling of decision-making regarding the management of its ventilation system under non-stationary wind loads. Fuzzy logic, proposed by L. Zadeh [14], allows for solving modelling and control tasks in various domains [12], and its combination with artificial neural networks has simplified the model-building processes [3, 13].

1. The aim and objectives of the study

The aim of this study is to develop a method for managing the ventilation system of the NSC based on neuro-fuzzy networks. To achieve this objective, it is necessary to analyze the current state of ventilation system management of the NSC, select approaches for automating the process, construct neuro-fuzzy networks that will model decision-making in ventilation control, and assess the quality of its performance on a test dataset.

2. Methods

Existing approaches to managing ventilation systems in the NSC are aimed at minimizing uncontrolled air leakage along with radioactive aerosols from its main space (MS) to the ring space (RS), and from there to the surrounding environment through the existing network of leaks. The leak area values were theoretically estimated during the design of the NSC at the beginning and end of its 100-year service life. However, after conducting hydraulic tests in 2017–2018, it was determined that the existing leak areas at that time already significantly exceeded their expected values at the end of the NSC's service life. There are forecasts that the values of existing leak areas will increase significantly over time [7]. Figure 1 shows a schematic representation of the conditional hydraulic airflow in the NSC [11].

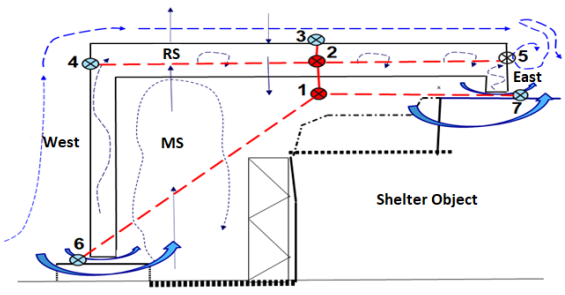


Fig. 1. Conditional hydraulic diagram of airflow movement in the NSC in a longitudinal section: 1 – MS node, 2 – RS node, 3 – node on the external cylindrical surface, 4 – node on the external surface of the western wall, 5 – node on the external surface of the eastern wall, 6 – node on the outer part of the western leaks, 7 – node on the outer part of the eastern releases [11]

As shown in Fig. 1, air masses enter the RS and MS of the NSC and are extracted from them by the ventilation system located on the cylindrical surface of the NSC, its eastern and western walls, as well as through the existing leaks, the size of which is not precisely determined.

Personnel managing the NSC ventilation system make decisions regarding the determination of air flow rates based on current data on climatic conditions (wind direction and speed), pressures in the MS and RS, and fan flow rates in the MS and RS. The creation of a digital twin of the NSC involves developing a control model for its ventilation system. Given the complexity of the process and the presence of uncertainty, it is advisable to apply intelligent information technologies based on fuzzy logic.

The existing NSC management information system contains historical data on decisions made regarding the management of ventilation systems, thus the use of neuro-fuzzy networks can be considered. In these networks, the inference process is based on fuzzy logic methods, while the construction of corresponding membership functions is achieved using neural network training methods, which simplifies the model development. The using of neuro-fuzzy networks allows for the consideration of the non-stationarity of the NSC, which is manifested in the changing leak areas. The network structure enables model adaptation, as it represents the system of fuzzy logical inference in the form of a neural network that is convenient for training, updating, and analysis.

Considering the results of comparative analysis of various types of neuro-fuzzy systems, the availability of information technologies for their implementation, the Adaptive Neuro-Fuzzy Inference System (ANFIS) proposed by R. Jang was chosen for modeling the control of NSC ventilation systems, implemented using MATLAB [3, 13]. The ANFIS network realizes the Takagi-Sugeno fuzzy inference system in the form of a five-layer feedforward neural network. The first layer of the network contains neurons corresponding to the terms of input linguistic variables. Gaussian functions were chosen as membership functions for the terms based on the results of fuzzy clustering of statistical data on decision-making in ventilation system management. The second layer consists of neurons that process the output values of the neurons in the first layer, forming the antecedents of rules. Each neuron in the second layer determines the degree of fulfillment of the corresponding rule in the knowledge base. The third layer consists of neurons that calculate normalized degrees of fulfillment of rules, and the fourth layer computes the rule consequents. The fifth layer contains only one neuron that calculates the final result of the network's operation.

One of the most important components of a fuzzy logic-based control system is the knowledge base, which represents a collection of fuzzy "if-then" rules that define the relationship between the inputs and outputs of the system under investigation. In this work, a Takagi-Sugeno rule base with linear functions as consequents was used for application in the ANFIS system, which has the following general form:

$$R_i: IF\ x_1\ is\ A_1\ AND\ x_2\ is\ A_2 \dots x_n\ is\ A_n\ THEN \quad (1)$$
$$y = b_{i1}x_1 + \cdots + b_{in}x_n + b_{i0}$$

where R_i is a rule of the number i of the fuzzy logic inference knowledge base, $x_1 \dots x_n$ are input variables characterizing of the state of the controlled object, $A_1 \dots A_n$ are terms of input linguistic variables corresponding to fuzzy variables with specific membership functions $\mu_1(x_1) \dots \mu_n(x_n)$, through which the expression $x_k\ is\ A_k$ is calculated as $\mu_k(x_k)$, y is the output variable determining the control of the controlled object and $b_{i0}, b_{i1} \dots b_{in}$ are parameters of a linear function that determines the value of the control variable according to the rule R_i.

The main indicators determined by decision-makers in managing the NSC ventilation system are air consumption for RS and MS. Considering that ANFIS allows the construction of fuzzy logic systems only for a single output variable, it is necessary to develop two separate models for calculating air consumption for RS and MS. Figure 2 shows the structure of the fuzzy logic inference system for controlling RS's air consumption, built in the ANFIS editor of MATLAB.

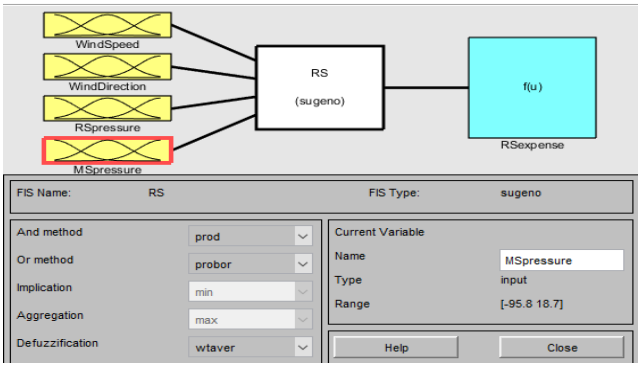


Fig. 2. Structure of the fuzzy logic inference system for controlling RS's air consumption

As seen in Fig. 2, the air consumption of the RS's ventilation system is determined by the wind speed and direction, as well as pressure differentials between RS and the surrounding environment. During the training of ANFIS using 25,170 records of statistical data, 162 fuzzy logic inference rules were generated to calculate RS's air consumption based on the Takagi-Sugeno method. The results of the neuro-fuzzy model training are presented in Fig. 3.

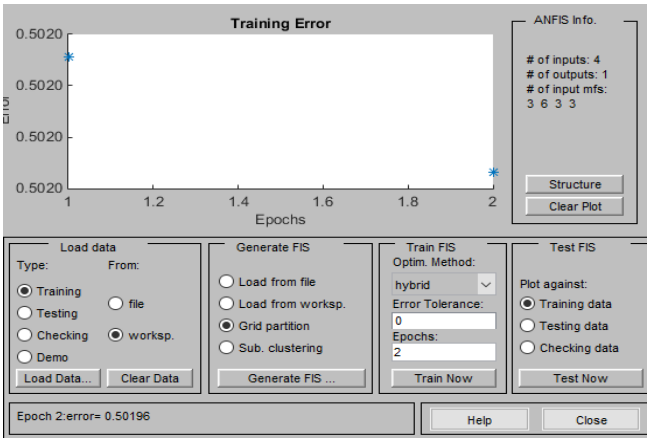


Fig. 3. The results of training the neuro-fuzzy network for controlling RS's air consumption

As shown in Fig. 3, the training was performed using a hybrid optimization method, which yielded a mean square accuracy of 0.50196, corresponding to a relative error of approximately 3.7%. Examples of plots depicting the membership functions of the input linguistic variables and the formed rule base of the fuzzy logic inference system are presented in Fig. 4 and 5, respectively. Based on the results of membership function formation using fuzzy clustering, Gaussian functions were selected as the basic type of membership functions. The number of linguistic variable terms was determined experimentally, aiming to achieve good accuracy while maintaining acceptable computation speed during the system's training. In accordance with the general form of the knowledge base rules (1), fuzzy logical conjunction is used in the rule conditions.

Figure 6 shows the graph depicting the relationship between the air consumption of the RS's ventilation system and the wind speed and direction.

As seen in Fig. 6, when the wind direction angle decreases, the air consumption of the RS's ventilation system exhibits a significantly nonlinear dependence on wind speed.

The accuracy evaluation of the trained neuro-fuzzy inference system on the test dataset containing 2517 records showed an error of 0.43023, which is approximately equal to a relative error of 3.2%. This indicates the system's adequacy and its practical potential for controlling the air consumption of the RS's ventilation system. The results of the accuracy evaluation of the neuro-fuzzy inference system are presented in Fig. 7.

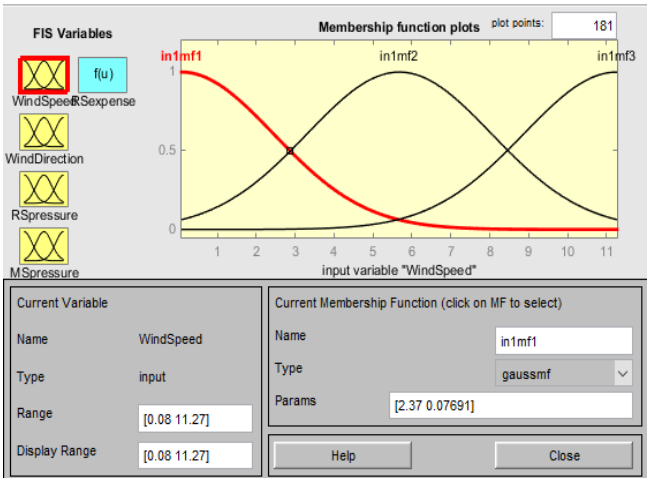


Fig. 4. Membership functions of the input linguistic variable "Wind Speed"

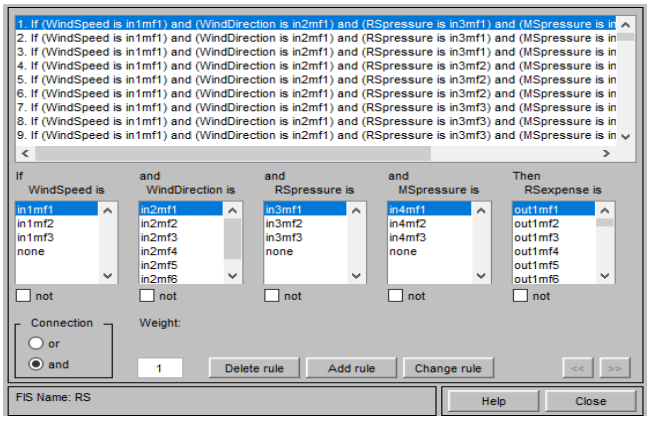


Fig. 5. The rule base for fuzzy control of air consumption by the RS's ventilation system

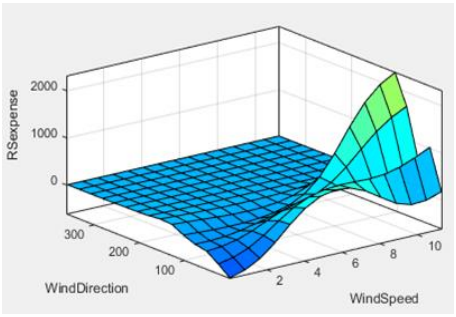


Fig. 6. Dependence of air consumption of the RS's ventilation system on wind speed and direction

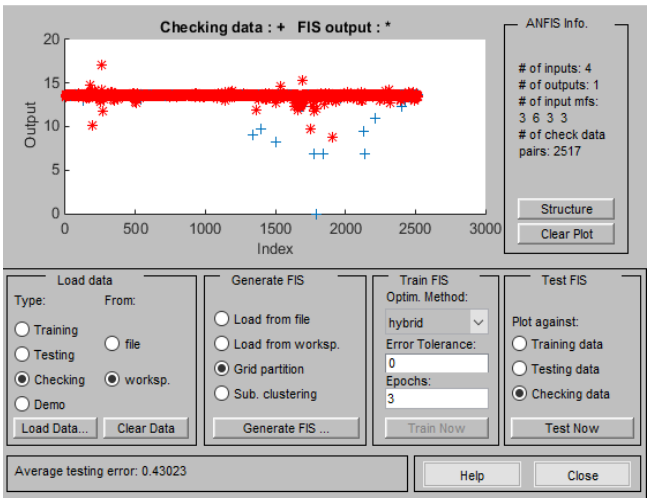


Fig. 7. The results of accuracy evaluation of the neuro-fuzzy network for air consumption of the RS's ventilation system on the test dataset

The control of air ventilation systems of the MS is carried out based on the same input variables as the control of the ventilation system of the RS. The results of training the ANFIS system for controlling the ventilation of the MS based on statistical data are shown in Fig. 8.

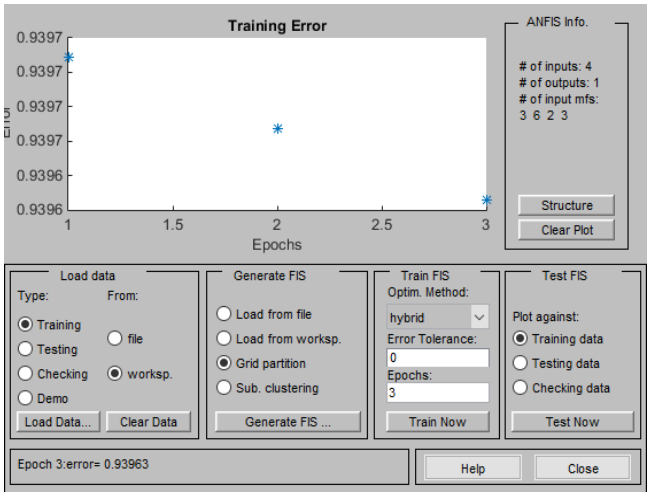


Fig. 8. The results of training the neuro-fuzzy network for controlling MS's air consumption

As shown in Fig. 8, the training accuracy reached 0.93963, which corresponds to approximately a 2.7% relative error on the training data. The knowledge base consists of 108 fuzzy production rules, enabling the calculation of ventilation system MS costs using the Takagi-Sugeno method. The results of the accuracy verification of the neuro-fuzzy network for controlling the expenses of MS's ventilation systems are presented in Fig. 9.

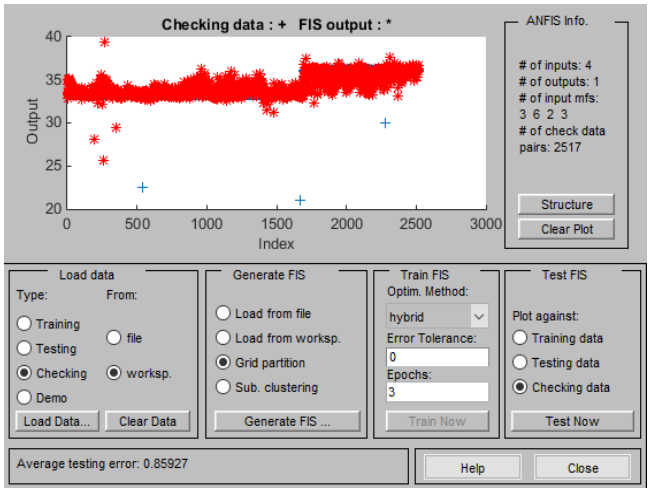


Fig. 9. The results of accuracy evaluation of the neuro-fuzzy network for air consumption of the MS's ventilation system on the test dataset

As observed in Fig. 9, a comparison of the neuro-fuzzy logic system's calculation results with the control dataset, consisting of 2517 records, revealed a deviation of 0.85927 on identical sets of input data, which approximately corresponds to a 2.5% relative error. The high accuracy indicators indicate the practical applicability of this system for managing the costs of MS's ventilation systems.

Based on the developed and trained neuro-fuzzy networks, a ventilation system control method NSC has been devised. The schematic representation of this method is shown in Fig. 10.

As shown in Fig. 10, based on processed sensor readings, input variable values are formed for the developed neuro-fuzzy networks, and using the Takagi-Sugeno method, the expenses of the ventilation systems KP and OO are calculated. These recommended expenses are analyzed and verified by decision-makers (DM). If the ventilation expenses are confirmed, they are transferred as tasks to the NSC's control system.

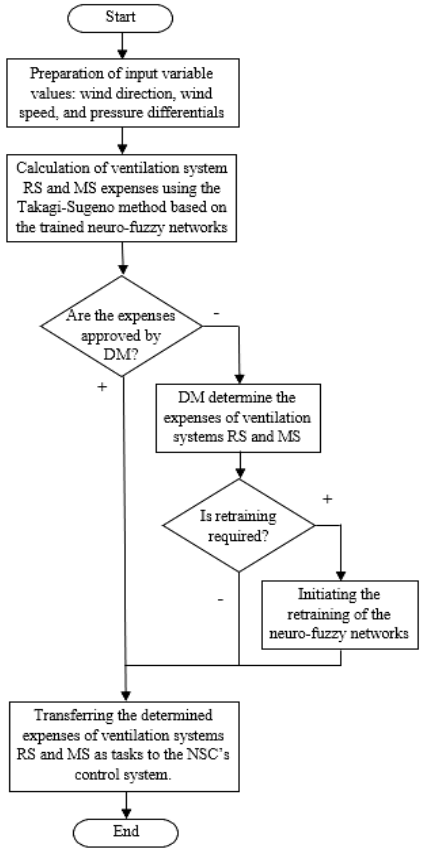


Fig. 10. The diagram of the neuro-fuzzy control method for the NSC's ventilation systems

Otherwise, the DM determines the ventilation expenses independently, after which the need for retraining the neuro-fuzzy models is evaluated. It is proposed to conduct retraining of these models if the deviations between the proposed system and DM values are significant or if the number of deviated decisions exceeds a certain threshold. The DM initiates the retraining of the models, and after making decisions, the results are transferred to the NSC's control system.

3. Conclusions

1) Research on the current state of ventilation management in NSC has been conducted. Decisions regarding the airflow volumes of the ventilation systems are made by operators based on current data on the state of the NSC and wind loads under conditions of uncertainty about existing leak areas. To automate the ventilation management, neuro-fuzzy modeling has been chosen.
2) Using the adaptive neuro-fuzzy inference system (ANFIS) and statistical data on NSC operations, neuro-fuzzy models have been developed, enabling the calculation of its ventilation system expenses using the Takagi-Sugeno method. Validation of the proposed approaches on a test dataset demonstrated sufficiently high accuracy in the calculations, confirming their practical applicability for making decisions regarding NSC's ventilation management.
3) Based on the developed and trained neuro-fuzzy networks, a ventilation system management method for the NSC has been devised. The method involves calculating ventilation system costs using neuro-fuzzy networks and approving or adjusting them by decision-makers.
4) The results of this work can be valuable for the automation and modeling of NSC processes using information technologies, including the development of digital twins of the NSC for process management and personnel training purposes.

References

[1] Batiy V. G. et al.: Dynamics of radioactive aerosol concentration during the removal of fuel-containing materials from the Shelter object. Nuclear and Radiation Safet 4, 2015, 41–44.
[2] Grieves M.: Virtually Intelligent Product Systems: Digital and Physical Twins. Complex Systems Engineering: Theory and Practice, 2019, 175–200.
[3] Jang J. S. R.: ANFIS: Adaptive-Network-Based Fuzzy Inference System. IEEE Transactions on Systems, Man, and Cybernetics 23(3), 1993, 665–685.
[4] Krasnov V. O. et al.: Shelter Object: 30 years after the accident. Institute for Problems of Nuclear Power Plants Safety, Chornobyl 2016.
[5] Kratz B., Wieduwilt F., Saveliev M.: Pillars for Establishing a Durable and Future-Proof IT Architecture Maturing Along with the NSC: Approaches from Continuous Integration to Service Mesh. Mathematical Modeling and Simulation of Systems 344, 2022, 43–57.
[6] Krukovsky P. G. et al.: CFD-model as a digital twin of the radiation state of the new safe confinement of the Chernobyl Nuclear Power Plant. Problems of Atomic Science and Technology (VANT) 4 (128), 2020, 54–62.
[7] Krukovskyi P. G. et al.: Unorganized emissions of air with radioactive aerosols from the new safe confinement of the Chernobyl Nuclear Power Plant into the surrounding environment. Issues of atomic science and technology 6, 2021, 181–186.
[8] Krukovsky P. G. et al.: Model of thermal, gas-dynamic, moisture, and radiation conditions of the new safe confinement and the Shelter Object. 2nd International Conference "Issues of Decommissioning of Nuclear Facilities and Environmental Recovery" INUDECO, Slavutych, 2017, 347–350.
[9] Krukovsky P. G. et al.: New safe confinement of the Chernobyl nuclear power plant (calculation-experimental analysis in design and operation). LLC "Franko Pak", Kyiv 2019.
[10] Loboda, P. P., Starovit, I. S.: Architecture of the digital twin for the New Safe Confinement of the ChNPP. Connectivity 2, 2022, 22–26.
[11] Pysmennyy Y. et al.: Development of special mathematical software for controlling the ventilation units of the new safe confinement of the ChNPP. Nuclear & Radiation Safety 2(94), 2022, 35–43.
[12] Shushura O. M. et al.: Simulation of information security risks of availability of project documents based on fuzzy logic. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska – IAPGOS 12(3), 2022, 64–68.
[13] Viharos Z., Kis K.: Survey on Neuro-Fuzzy systems and their applications in technical diagnostics and measurement. Measurement 67, 2015, 126–136.
[14] Zadeh L. A.: Fuzzy sets. Information and Control 8, 1965, 338–353.

M.Sc. Petro P. Loboda
e-mail: lobodapetr@gmail.com

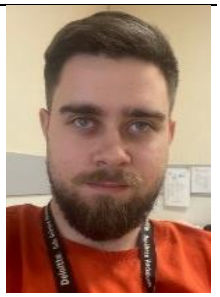
Master of computer science, Ph.D. student, Department of Software Engineering in Energy, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine. Author of 9 scientific publications.

http://orcid.org/0000-0002-0808-8014

M.Sc. Ivan S. Starovit
e-mail: ivanstarovit@gmail.com

Master of Engineering Sciences, Ph.D. student, Department of Software Engineering in Energy, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine. Author of 8 scientific articles in professional journals, of which 2 are in scientometric databases Scopus and Web of Science.

http://orcid.org/0000-0002-6511-4610



D.Sc. Oleksii M. Shushura
e-mail: leshu@i.ua

Doctor of Technical Sciences, associate professor, professor of the Department of Digital Technologies in Energy, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine. Author of over 117 publications, including 1 monographs, 5 textbooks, more than 48 scientific articles in professional journals, of which 2 are in the scientometric databases Scopus and Web of Science.

http://orcid.org/0000-0003-3200-720X

Prof. Yevhen V. Havrylko
e-mail: gev.1964@ukr.net

Doctor of Engineering Sciences, professor, Department of Software Engineering in Energy, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine. Author of more than 50 publications, including 3 textbooks, 3 patents for inventions and more than 35 scientific articles in professional journals, of which 17 are in scientometric databases Scopus and Web of Science.

http://orcid.org/0000-0001-9437-3964

Ph.D. Maxim V. Saveliev
e-mail: m.saveliev@ispnpp.kiev.ua

Over 25+ years of international experience in computerized systems in Nuclear Industry. As a senior researcher at the Institute for Safety Problems of Nuclear Power Plants of Ukraine National Academy of Science, he performs scientific monitoring of data from destroyed Reactor No 4 of Chernobyl, and conducts research on Robotics applications for Nuclear Industry, Digital Twining, and Nuclear Waste Management. Author of 54 scientific publications.

http://orcid.org/0000-0002-2118-4748

Ph.D. Eng. Natalia Sachaniuk-Kavets'ka
e-mail: skn1901@gmail.com

In 2003 she received a Candidate of Sciences degree (Ph.D.) at the State Research Institute of Information Infrastructure of the State Committee for Communications and Informatization of Ukraine and the National Academy of Sciences of Ukraine (Lviv, Ukraine) with a specialty in automated control systems and advanced information technologies. Currently, she is an associate professor at the Department of Higher Mathematics of the Vinnytsia National Technical University (Vinnytsia, Ukraine). Scientific research includes issues related to mathematical modeling of information processing in a logical-temporal environment, as well as the possibility of using mathematical models of image processing for information protection.

http://orcid.org/0000-0001-6405-1331

Ph.D. Oleksandr Neprytskyi
e-mail: neprytskyi@gmail.com

Associate professor at the Vinnytsia Mykhailo Kotsiubynskyi State Pedagogical University.
Scientific research includes issues related to mathematical modeling of information processing in society.

http://orcid.org/0000-0001-9536-2538

Ph.D. Dina Oralbekova
e-mail: z.kalpeyeva@satbayev.university

Satbayev University, Almaty, Kazakhstan,
Ph.D. in 2022 completed the full course of the doctoral program in the direction of EP 8D06103 – "Management information systems" at Satbaev University and works as a senior researcher at the IICT.
Research interests: machine learning, natural language processing, technologies for the development and processing of speech recognition systems.

http://orcid.org/0000-0003-4975-6493

M.Sc. Dinara Mussayeva
e-mail: d_i_n_mus@mail.ru

A master of Al-Farabi Kazakh National University, Ph.D. student of the Institute of Economics of the CS MES RK, specialty 6D051300 – "World Economy". She started her career at the Institute of Economics as a junior researcher. Currently, she is the Scientific Secretary and Research Associate of the IE. As a responsible executive, she actively participates in conducting fundamental and applied research in priority areas for the state, such as digitalization of the economy, knowledge-based economy, etc.

http://orcid.org/0000-0002-8349-213X

http://doi.org/10.35784/iapgos.5302 received: 14.09.2023 | revised: 29.10.2023 | accepted: 30.11.2023 | available online: 20.12.2023

MODEL OF THE FLAT FAIRING ANTENNA DIELECTRIC LAYER WITH AERODYNAMIC HEATING

Valeriy Kozlovskiy[1], Valeriy Kozlovskiy[2], Oleksii Nimych[2], Lyudmila Klobukova[2], Nataliia Yakymchuk[3]
[1]National Technical University of Ukraine "Kyiv Polytechnic Institute named after Igor Sikorsky", Institute of Special Communications and Information Protection, Kyiv, Ukraine, [2]National Aviation University, Faculty of Cybersecurity, Computer and Software Engineering, Kyiv, Ukraine, [3]Lutsk National Technical University, Faculty of Computer and Information Technologies, Lutsk, Ukraine

Abstract. *To protect the antenna systems of modern aircraft, radio-transparent dielectric fairings are widely used. At low flight speeds, when designing and evaluating the characteristics of the fairing-antenna, it is assumed that the dielectric constant is a constant value and does not depend on the aircraft's flight speed. As the flight speed increases, as a result of aerodynamic heating of the fairing, its dielectric permeability changes, which leads to errors in the processing of received signals. Currently, to take into account the effect of dielectric coatings heating when designing antenna systems, the temperature of the fairing wall is averaged over its thickness. This method during maneuvering and at high flight speeds leads to large errors in determining the characteristics of the fairing antenna since the nature of the temperature distribution along the thickness of the fairing wall is not taken into account. A new approach to the analysis of dielectric layers with their uneven heating along the thickness is proposed. The obtained results make it possible to adjust the signal processing algorithms with analog and digital matrices, as a result of taking into account the emerging heat flows affecting the fairing of the aviation antenna, which leads to the improvement of the characteristics of the antenna systems.*

Keywords: aviation antenna, dielectric layer, aerodynamic heating, wave resistance, quadrupole

MODEL PŁASKIEJ WARSTWY DIELEKTRYCZNEJ ANTENY Z NAGRZEWANIEM AERODYNAMICZNYM

Streszczenie. *Aby chronić systemy antenowe nowoczesnych samolotów, szeroko stosuje się radioprzepuszczalne owiewki dielektryczne. Przy małych prędkościach lotu przy projektowaniu i ocenie charakterystyk anteny owiewkowej przyjmuje się, że stała dielektryczna jest wartością stałą i nie zależy od prędkości lotu samolotu. Wraz ze wzrostem prędkości lotu, w wyniku nagrzewania się aerodynamicznego owiewki, zmienia się jej przepuszczalność dielektryczna, co prowadzi do błędów w przetwarzaniu odbieranych sygnałów. Obecnie, aby uwzględnić wpływ nagrzewania powłok dielektrycznych przy projektowaniu systemów antenowych, temperaturę ścianki owiewki uśrednia się w stosunku do jej grubości. Metoda ta podczas manewrowania i przy dużych prędkościach lotu prowadzi do dużych błędów w określaniu charakterystyk anteny owiewki, gdyż nie bierze się pod uwagę charakteru rozkładu temperatury wzdłuż grubości ścianki owiewki. Zaproponowano nowe podejście do analizy warstw dielektrycznych przy ich nierównomiernym nagrzewaniu na całej grubości. Uzyskane wyniki pozwalają na dostosowanie algorytmów przetwarzania sygnału z matrycami analogowymi i cyfrowymi, w wyniku uwzględnienia powstających strumieni ciepła wpływających na owiewkę anteny lotniczej, co prowadzi do poprawy charakterystyki systemów antenowych.*

Słowa kluczowe: antena lotnicza, warstwa dielektryczna, nagrzewanie aerodynamiczne, opór falowy, kwadrupol

Introduction

At high flight speeds of aircraft (LA), aerodynamic heating of radio-transparent fairings (coatings) occurs. The reason for this is the influence of friction due to the viscosity of the air and the roughness of the fairing surface. The most intense heating of the material of the walls of the fairing leads to a change in the dielectric constant and losses, which is the reason for the change in the conditions for the passage of radio waves. As a result, the characteristics of the radome-antenna system differ from the calculated ones, which is an additional source of errors in the processing of received signals. The issue of taking into account this class of errors is especially acute in digital signal processing, when, during various types of aircraft maneuvers, the flight speed and altitude change, which leads to a change in the temperature of the fairing heating in time, as a result of which the characteristics of the fairing-antenna system also change in time. Thus, the problem of correcting the digital signal processing algorithm with an antenna array arises. For example, at different temperatures of heating the fairing wall above the elements of the antenna array, the delay time of the received signals changes, which leads to an increase in the direction finding errors of the phase direction finders. Obviously, with a change in the heating temperature, the value of the direction-finding error will also change.

The goal is to define a four-pole model of a flat dielectric layer during its aerodynamic heating. Representation of the layer in the form of a four-terminal network makes it possible to take into account the effect of temperature distribution on the passage of an electromagnetic field through a dielectric layer under various boundary conditions.

The paper poses the problem of finding the electro-magnetic field in a flat dielectric layer when a plane electromagnetic wave falls on its surface under conditions of aerodynamic heating. To solve the problem, it is first necessary to determine the distribution of temperature and permittivity over the thickness of the dielectric layer. Further, using the permittivity, we find the dependence of the wave resistance on the thickness and using the wave resistance, we determine the parameters of the equivalent quadrupole. The system of Z-parameters (resistance matrix) was chosen as the system of parameters of the quadrupole.

1. Literature review

In [1–25] are devoted to the issues of determining the electromagnetic field in the walls of the antenna radome and the creation of special radome materials, from which it follows that all materials used in the construction of microwave radomes can be divided into two groups [1, 5, 8, 10–12, 16, 20, 22, 23, 25, 26]: 1) structural materials that do not have an interaction device with an electromagnetic field during operation; 2) materials interacting with an electromagnetic field during aerodynamic heating, including with a laser. If there are no requirements for materials of the first group with their properties in the microwave, then for materials of the second group these requirements are mandatory, especially when it comes to millimeter microwaves and high-power pulsed lasers [2–4, 6, 7, 9, 13–15, 17–19, 21, 24, 25].

In the vast group of microwave materials, a special place is occupied by solid dielectrics, which are both structural materials that act as mechanical carriers of the structure and microwave materials operating in electromagnetic fields [2, 5, 7–9, 11–14, 17–20, 22–26].

Almost always, the most stringent requirements for the radio engineering parameters of dielectrics are imposed in connection with the special conditions of their operation. In particular, it should be noted the high requirements for the stability of the parameters and characteristics of dielectrics subjected to temperature effects. The properties of most of the known microwave dielectrics significantly depend on temperature, which ultimately changes the operational characteristics of radio engineering devices as a whole. This stimulates the development

of new heat-resistant and thermally stable dielectric materials [14, 17, 24] and the development of new methods and specialized research equipment for measuring at ultrahigh frequencies the temperature dependencies of material properties under thermophysical conditions similar to operational ones. Known methods for determining the temperature dependencies of the parameters of microwave dielectrics: waveguides, resonators, optical, and some others do not always solve the problem properly (there are unacceptably large measurement errors).

Existing methods [7, 8, 10–12, 16, 20, 24, 25] do not provide information about the dynamic dependence of the parameters of dielectric microwave materials on temperature during heating. It should be emphasized that the problem of measuring the parameters of dielectric materials when their physical state changes with simultaneous exposure to laser and microwave radiation have not been solved in modern microwave measuring technology. Research of this kind is often necessary for the theory and practice of designing various devices operating under conditions of aerodynamic heating.

At present, when designing fairings, aerodynamic heating is taken into account by averaging the temperature over the wall thickness [8, 22, 25], and the permittivity corresponding to the average temperature is used in calculations. The averaging method makes it possible to obtain satisfactory results under stationary flight conditions, when the heating temperature is constant throughout the thickness of the coating, that is, in the absence of a temperature gradient. When performing aircraft maneuvers, the conditions for heating the fairing change in time and wall thickness. In this case, the averaging method leads to large errors, since the change in the temperature gradient over time is not taken into account.

2. Researches methodology

When calculating electrodynamic objects, methods based on well-known solutions for a site of sufficiently small dimensions with a uniform distribution of the electric (E) and magnetic (H) fields are widely used. In this case, such an area can be replaced by a flat layer with a constant value of the vectors E and H. This layer, as an element of the fairing, is subjected to aerodynamic heating with an uneven temperature distribution over the thickness (figure 1), where T_1 is the temperature of the outer surface of the layer from the side of the heat flow, K; T_2 is the temperature of the inner surface of the layer, K. As the material of the layer, we choose quartz ceramics, which is widely used in the creation of fairings. It is known that the dependence of the absolute permittivity of quartz ceramics depends on temperature [10, 16, 20, 25].

$$\varepsilon(x) = \varepsilon_0 exp\left[1 - P + 2.6\cdot10^{-5}\left(T(x) - 290\right)\right] \tag{1}$$

where $T(x)$ is the temperature distribution, K; x – thickness of the dielectric layer; P – porosity (volume fraction of pores); ε_0 – absolute permittivity at zero porosity and T = 290 K. Numerical coefficient $2.6\cdot10^{-5}$ does not play a fundamental role and is determined by the type of material used.

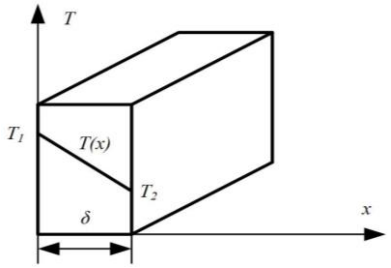


Fig. 1. Temperature distribution over the thickness of the dielectric layer

According to the law of temperature distribution over the layer thickness [6], we have

$$T(x) = (T_2 - T_1)\frac{x}{\delta} + T_1 \tag{2}$$

Substitute (2) into (1):

$$\begin{aligned}\varepsilon(x) = \varepsilon_0 \exp\left[1 - P + 2.6\cdot10^{-5}\left(T(x) - 290\right)\right] = \\ \varepsilon_0 \exp\left[1 - P + 2.6\cdot10^{-5}\left(T_1 - 290\right)\right]\cdot \\ \cdot\exp\left[2.6\cdot10^{-5}\left((T_2 - T_1)\frac{x}{\delta}\right)\right]\end{aligned} \tag{3}$$

Let us introduce the following notation:

$$A = 2.6\cdot10^{-5} \tag{4}$$

$$\begin{aligned}A_1 = 1 - P + 2.6\cdot10^{-5}\left(T_1 - 290\right) = \\ 1 - P + A\left(T_1 - 290\right)\end{aligned} \tag{5}$$

As a result, the permittivity can be written as:

$$\varepsilon(x) = \varepsilon_0 exp(A_1)\cdot exp\left[A\left((T_2 - T_1)\frac{x}{\delta}\right)\right] \tag{6}$$

Non-magnetic materials with an absolute magnetic permeability μ equal to the absolute magnetic permeability of a vacuum are usually used as dielectric coatings for aircraft antennas. In this case, the wave resistance of the layer at the normal incidence of a plane transverse wave, taking into account (6), is determined by the expression

$$\begin{aligned}W(x) = \sqrt{\frac{\mu}{\varepsilon(x)}} = \sqrt{\frac{\mu}{\varepsilon_0 exp(A_1)\cdot exp\left[A\left((T_2 - T_1)\frac{x}{\delta}\right)\right]}} = \\ = \sqrt{B}\exp\left[A\left((T_1 - T_2)\frac{x}{2\delta}\right)\right]\end{aligned} \tag{7}$$

Value

$$B = \frac{\mu}{\varepsilon_0 exp(A_1)} \tag{8}$$

Denote

$$2\alpha = A\left((T_1 - T_2)\frac{1}{2\delta}\right) = 2.6\cdot10^{-5}\left((T_1 - T_2)\frac{1}{2\delta}\right) \tag{9}$$

Then the wave resistance of the layer and the permittivity

$$W(x) = \sqrt{B}e^{2\alpha x},\ \varepsilon(x) = \varepsilon_0 e^{A_1}\cdot e^{-4\alpha x} \tag{10}$$

It follows from the analysis of expression (10) that the temperature difference on the outer and inner walls strongly affects the distribution of wave resistance over the layer thickness (figure 2).

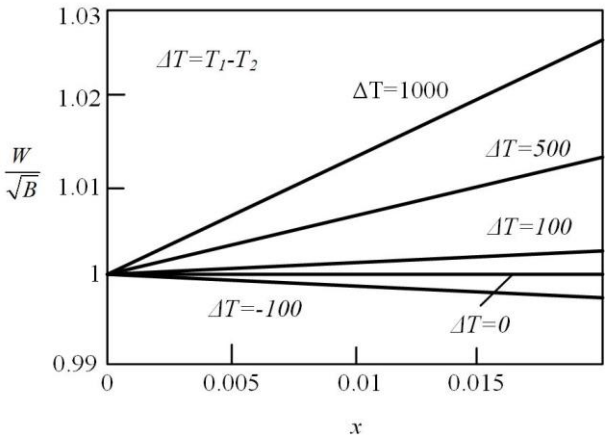


Fig. 2. Dependence of the wave resistance on the thickness of the dielectric layer with a total thickness of δ = 0.02 m

In the general case, processes in flat coatings under the incidence of an electromagnetic field that varies in time according to the harmonic law $e^{j\omega t}$, are described by the Maxwell equations:

$$rot\bar{H} = j\omega\varepsilon\bar{E},\ \ rot\bar{E} = j\omega\mu\bar{H} \tag{11}$$

where ε and μ are the absolute permittivity and permeability of the medium.

Let us assume that the parameters of the coverage environment change only along the z coordinate: $\varepsilon = \varepsilon(z), \mu = \mu(z)$. Consider the passage of a plane electro–magnetic wave (PEW) through an inhomogeneous layered medium. For definiteness, we assume that a plane wave has components E_x, H_y. Then equations (11) are transformed to the form

$$-\frac{dE_x}{dz} = j\omega\mu H_y,\quad -\frac{dH_y}{dz} = j\omega\varepsilon E_x \tag{12}$$

Let's introduce a new variable

$$\tau = \int_0^z \sqrt{\varepsilon(x)\mu(x)} \tag{13}$$

Obviously, τ is the propagation time of the plane wave from the beginning of the layer $(z = 0)$ to the points of the layer with the z coordinate, that is, the variable τ is the delay time of a layer of thickness z. In this case, the system of equations (12) takes the form:

$$-\frac{dE}{d\tau} = W(\tau)\, j\omega H,\quad -\frac{dH}{d\tau} = W^{-1}(\tau)\, j\omega E \tag{14}$$

where $W(\tau) = \sqrt{\mu(\tau)/\varepsilon(\tau)}$ is the wave impedance depending on the delay time τ, indices for E and H are omitted for simplicity.

Differentiating equations (14), we obtain equations for the strengths of the electric and magnetic fields E and H:

$$\mathrm{E}''(\tau) - \frac{\mathrm{W}'(\tau)}{\mathrm{W}(\tau)} E'(\tau) - p^2 E(\tau) = 0 \tag{15}$$

$$\mathrm{H}''(\tau) + \frac{\mathrm{W}'(\tau)}{\mathrm{W}(\tau)} H'(\tau) - p^2 H(\tau) = 0 \tag{16}$$

where $p = j\omega$ is a complex frequency variable.

Let's find the relationship between the delay time τ and layer thickness x. To do this, we substitute (6) with (13). In this case, since the heat flux propagates along the x-axis (figure 1), for convenience, we will assume that the layer is also inhomogeneous along the x-axis, that is, the PEW wave propagates along the x-axis. In this case, all formulas remain unchanged. The only components of the PEW field are E_y, H_z. Taking into account the above, we have the dependence of the delay time on the thickness (figure 3):

$$\begin{aligned}\tau &= \int_0^x \sqrt{\mu\varepsilon(\xi)}d\xi = \int_0^x \sqrt{\mu}\cdot\sqrt{\varepsilon_0\exp(\mathrm{A}_1)\cdot\exp(-4\alpha\xi)}d\xi = \\ &= \frac{b}{2\alpha}\left(1 - e^{-2\alpha x}\right),\ \ \mathrm{b} = \sqrt{\mu\varepsilon_0\exp(\mathrm{A}_1)}\end{aligned} \tag{17}$$

From formula (17) it follows that the delay time of the layer with thickness δ (figure 1) is equal to

$$\mathrm{t}_3 = \frac{b}{2\alpha}\left(1 - e^{-2\alpha\delta}\right) \tag{18}$$

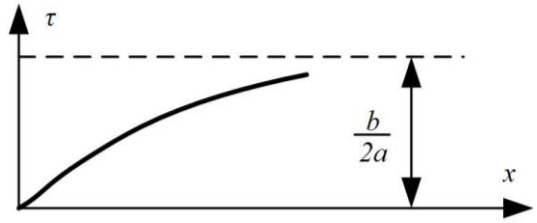


Fig. 3. Dependence of the delay time on the layer thickness

Using (17), we express the wave resistance $W(x)$ in terms of $W(\tau)$. For this, from expression (17) we find

$$-2\alpha x = \ln\left(1 - \tau\frac{2\alpha}{\mathrm{b}}\right) \tag{19}$$

Whence it follows that the wave resistance is determined by the expression

$$W(\tau) = \sqrt{B}e^{2\alpha x} = \sqrt{B}\frac{1}{1 - \dfrac{2\alpha}{b}\tau} \tag{20}$$

To determine the distribution of the electromagnetic field, we use equations (15), and (16). First, we find a solution for the electric field strength. To do this, we rewrite equation (15) in the form

$$\mathrm{E}''(\tau) - \frac{\mathrm{W}'(\tau)}{\mathrm{W}(\tau)} E'(\tau) + \omega^2 E(\tau) = 0 \tag{21}$$

According to (20), we find

$$\mathrm{W}'(\tau) = \frac{\mathrm{dW}(\tau)}{\mathrm{d}\tau} = \sqrt{B}\frac{\dfrac{2\alpha}{b}}{\left(1 - \dfrac{2\alpha}{b}\tau\right)^2} \tag{22}$$

where

$$\frac{\mathrm{W}'(\tau)}{\mathrm{W}(\tau)} = \frac{\dfrac{2\alpha}{b}}{1 - \dfrac{2\alpha}{b}\tau} = \frac{1}{\dfrac{b}{2\alpha} - \tau} \tag{23}$$

Substituting (23) into (21), we obtain the equation

$$\mathrm{E}''(\tau) + \frac{1}{\tau - \dfrac{\mathrm{b}}{2\alpha}} E'(\tau) + \omega^2 E(\tau) = 0 \tag{24}$$

Let us transform equation (24). To this end, we introduce a new variable

$$y = \tau - \frac{b}{2\alpha} \tag{25}$$

In this case

$$\frac{dE}{d\tau} = \frac{dE}{dy}\frac{dy}{d\tau} = \frac{dE}{dy},\quad \frac{d^2E}{d\tau^2} = \frac{d^2E}{dy^2}$$

and equation (21) takes the form

$$\mathrm{E}''(\mathrm{y}) + \frac{1}{y}E'(y) + \omega^2 E(y) = 0 \tag{26}$$

This equation is a modified Bessel equation, a particular solution of which is the zero-order Bessel function of the first kind $J_0(\omega y)$ [9]:

$$\frac{d^2 J_0(\omega y)}{dy^2} + \frac{1}{y}\frac{dJ_0(\omega y)}{dy} + \omega^2 J_0(\omega y) = 0 \tag{27}$$

The general solution to (26) is the function [9]

$$E(y) = C_1 J_0(\omega y) + C_2 Y_0(\omega y) \tag{28}$$

where $Y_0(\omega y)$ is the zero-order Bessel function of the second kind (the Neumann, Weber function). The constants C_1, C_2 are determined by the boundary conditions.

According to (14), the magnetic field strength is determined by the expression

$$H(\tau) = -\frac{E'(\tau)}{j\omega W(\tau)} \tag{29}$$

Returning to the variable τ, we rewrite (28)

$$\mathrm{E}(\tau) = C_1\omega J_0\left(\omega\left(\tau - \frac{\mathrm{b}}{2\alpha}\right)\right)C_2\omega Y_0\left(\omega\left(\tau - \frac{\mathrm{b}}{2\alpha}\right)\right) \tag{30}$$

To determine the magnetic field strength H, we use (29). Given that [25]

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}J_0\left(\omega\left(\tau - \frac{\mathrm{b}}{2\alpha}\right)\right) &= -\omega J_1\left(\omega\left(\tau - \frac{\mathrm{b}}{2\alpha}\right)\right) \\ \frac{\mathrm{d}}{\mathrm{d}\tau}\mathrm{Y}_0\left(\omega\left(\tau - \frac{\mathrm{b}}{2\alpha}\right)\right) &= -\omega Y_1\left(\omega\left(\tau - \frac{\mathrm{b}}{2\alpha}\right)\right)\end{aligned} \tag{31}$$

find

$$E'(\tau) = -C_1\omega J_1\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right) - C_2\omega Y_1\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right) \quad (32)$$

Therefore, according to (29)

$$H(\tau) = \left[C_1 J_1\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right) + C_2 Y_1\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right)\right]\cdot\frac{1-\frac{2a}{b}\tau}{j\sqrt{B}} \quad (33)$$

When analyzing a multilayer coating, it is convenient to represent it in the form of a cascade connection of four-terminal networks. In this case, each quadrupole is described by its system of parameters [18]: the system of Z-parameters, Y-parameters, the system of A-parameters (chain parameters), and various systems of wave parameters. At the same time, by knowing any system of parameters, it is possible to determine all the others. For definiteness, let's find the system of Z-parameters of one layer. Moreover, as Z-parameters, we take the surface resistance of the layer. In this case, according to (12...14), the layer can be represented as an inhomogeneous transmission line with a wave impedance $W(\tau)$ with voltage $U(\tau) = E(\tau)$ and current $I(\tau) = H(\tau)$. Let us agree that the potential of the upper terminal is greater than the potential of the lower terminal at the input and output of the quadrupole, and the input and output currents are directed towards the quadrupole. Then $E(\tau), H(\tau)$ there is a connection between [18]:

$$\begin{aligned} E(0) &= Z_{11}H(0) + Z_{12}H(t_з) \\ E(t_з) &= Z_{21}H(0) + Z_{22}H(t_з),\ Z_{12} = Z_{21} \end{aligned} \quad (34)$$

In general, the surface resistance of the layer, as follows from (30, 33), is equal to

$$Z_{in}(\tau) = \frac{C_1 J_0\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right) + C_2 Y_0\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right)}{\left[C_1 J_1\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right) + C_2 Y_1\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right)\right]\frac{1-\frac{2a}{b}\tau}{j\sqrt{B}}} \quad (35)$$

where the constants C_1, C_2 are determined by the surface resistance of the adjacent medium Z_a (the medium at the point $\tau = t_з$). Assuming $\tau = t_з$ we have

$$Z_a = \frac{C_1 J_0\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right) + C_2 Y_0\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right)}{\left[C_1 J_1\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right) + C_2 Y_1\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right)\right]\frac{1-\frac{2a}{b}t_з}{j\sqrt{B}}} \quad (36)$$

Let us express the input resistance $Z_{in}(\tau)$ in terms of the surface resistance of the adjacent medium Z_a, C_1, C_2. Then, according to (30, 33), we can write

$$\begin{aligned} &C_1 J_0\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right) + C_2 Y_0\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right) = E(t_з) \\ &\left[C_1 J_1\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right) + C_2 Y_1\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right)\right]\frac{1-\frac{2a}{b}t_з}{j\sqrt{B}} = H(t_з) \end{aligned} \quad (37)$$

From the system of equations (37) and expression (36), we find

$$C_1 = H_a\frac{Z_a Y_1 - \frac{Y_0}{F(t_з)}}{J_0Y_1 - J_1Y_0},\quad C_2 = H_a\frac{\frac{J_0}{F(t_з)} - Z_a J_1}{J_0Y_1 - J_1Y_0} \quad (38)$$

where the function argument J_0, J_1, Y_0, Y_1 is $\omega\left(t_з-\frac{b}{2\alpha}\right)$ (37).

Value $F(\tau)$ is defined by the expression

$$F(\tau) = \frac{1-\frac{2a}{b}\tau}{j\sqrt{B}} \quad (39)$$

Thus, the surface resistance (35) is transformed into the form

$$Z_{in}(\tau) = \frac{\left(Z_a Y_1 - \frac{Y_0}{F(t_з)}\right)J_0\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right) + \left(\frac{J_0}{F(t_з)} - Z_a J_1\right)Y_0\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right)}{\left[\left(Z_a Y_1 - \frac{Y_0}{F(t_з)}\right)J_1\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right) + \left(\frac{J_0}{F(t_з)} - Z_a J_1\right)Y_1\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right)\right]F(\tau)} \quad (40)$$

Using (40) we find the elements of the Z-parameter system (34). An element Z_{11} at an arbitrary value of the coordinate τ is determined by expression (40) at $Z_a = \infty$.

$$Z_{11}(\tau) = \frac{Y_1\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right)J_0\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right) - J_1\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right)Y_0\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right)}{\left[Y_1\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right)J_1\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right) - J_1\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right)Y_1\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right)\right]F(\tau)} \quad (41)$$

To define elements Z_{22}, Z_{12} we use (40). To do this, we find the layer resistance under the condition that $Z_a = 0$, $Z_a = R$, where R is a positive real number (resistance).

Then the short circuit layer resistance $Z_{sc}(Z_a = 0)$ $Z_R(Z_a = R)$ can be expressed in terms of the elements of the resistance matrix [18]:

$$Z_{sc} = Z_{11} - \frac{Z_{12}^2}{Z_{22}},\ Z_R = Z_{11} - \frac{Z_{12}^2}{Z_{22}+R} \quad (42)$$

From the system (42) we find the elements Z_{22}, Z_{12}:

$$Z_{22} = \frac{R(Z_{11}-Z_R)}{Z_R - Z_{sc}},\ Z_{12}^2 = \frac{R(Z_{11}-Z_R)(Z_{11}-Z_{sc})}{Z_R - Z_{sc}} \quad (43)$$

$$Z_{sc}(\tau) = \frac{-Y_0\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right)J_0\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right) + J_0\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right)Y_0\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right)}{\left[-Y_0\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right)J_1\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right) + J_0\left(\omega\left(t_з-\frac{b}{2\alpha}\right)\right)Y_1\left(\omega\left(\tau-\frac{b}{2\alpha}\right)\right)\right]F(\tau)} \quad (44)$$

Thus, expressions (41)–(44) completely determine the elements of the resistance matrix of a four-terminal network formed by a dielectric layer during aerodynamic heating.

3. Experiments

Based on the results obtained, the influence of aerodynamic heating of the radome of a two-element antenna array (AR), figure 4, on determining the angle of wave arrival was studied.

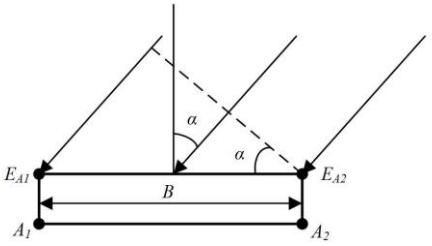


Fig. 4. Two-element antenna array under the dielectric layer

The absolute error in determining the direction of arrival of the wave that occurs when the aerodynamic heating of the fairing is not taken into account is determined by the expression

$$\Delta\phi = \phi_0 - \alpha = \arcsin\frac{(\Delta\Phi + \beta_1 - \beta_2)\lambda}{2\pi B} - \arcsin\frac{\Delta\Phi\lambda}{2\pi B},\ \Delta\Phi = \frac{2\pi B}{\lambda}\sin\alpha \quad (45)$$

where β_1, β_2 are the arguments of the dielectric transmission coefficient over the element A_1, A_2, respectively; ϕ_0 – the angle of arrival of the wave, calculated taking into account the heating of the fairing, α – the angle of arrival of the wave, calculated without taking into account the heating of the fairing.

P9606 pyroceramic (USA) with a relative permittivity at porosity $P = 0\%$ equal to $\varepsilon_r = 5.8$, oscillation frequency $f = 10$ GHz was used as the fairing material.

4. Results

On figures 5–8 show the dependences of the angular error of arrival of the TE wave on various factors.

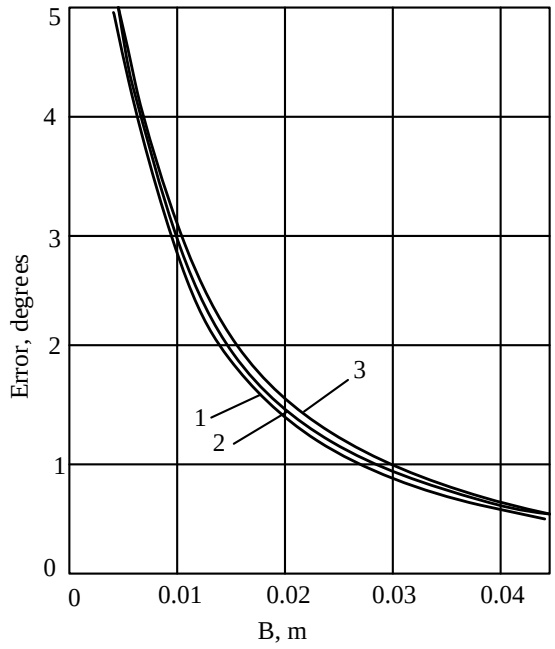


Fig. 5. Dependence of the angular error on the base (meter) at different angles of wave incidence: 1 – $\varphi_0 = \pi/16$, 2 – $\varphi_0 = \pi/8$; 3 – $\varphi_0 = \pi/3$; oscillation frequency $f = 10$ GHz; the heating temperature of the dielectric above the lattice element A_1 is $T_1 = 290$ K, and above the element A_2 – $T_2 = 1000$ K

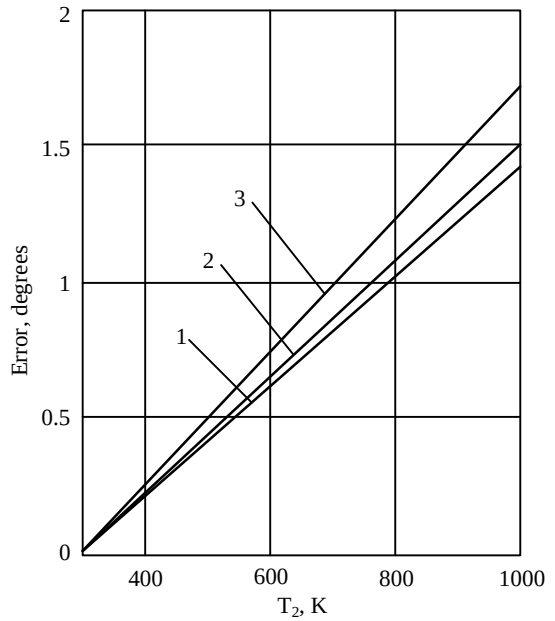


Fig. 6. Dependence of the angular error on temperature T_2 at $T_1 = 290$ K, $B = 0.015$ m, $P = 30\%$, material – P9606 pyroceramic, oscillation frequency $f = 10$ GHz: $\varphi_0 = \pi/16$; 2 – $\varphi_0 = \pi/4$; 3 – $\varphi_0 = \pi/3$

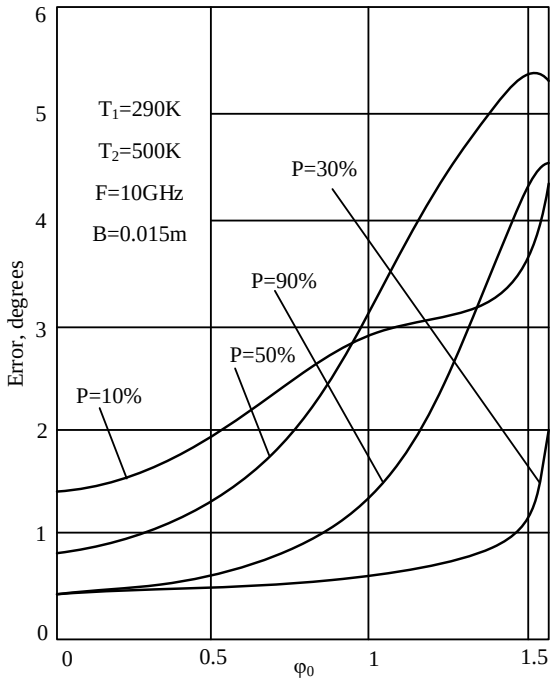


Fig. 7. Dependence of the angular error on the angle of incidence of the wave at different porosities P, material – P9606 pyroceramic

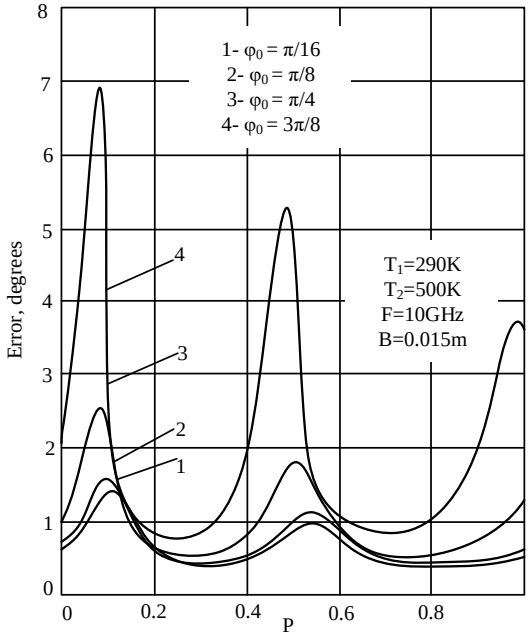


Fig. 8. Dependence of the angular error on porosity for P9606 pyroceramics at various angles of incidence φ_0

5. Discussion

It follows from the dependences in figures 5–8 that the angular error in determining the arrival of an electromagnetic wave depends not only on the size of the AA base, but also on the temperature of the aerodynamic heating of the dielectric wall of the fairing, the porosity of the dielectric, and the angle of arrival of the wave. From the analysis of the obtained dependencies, it follows:

1. With an increase in the temperature difference between heating the dielectric over various elements of the AR, the angular error increases according to a lawyer close to linear. In this case, with an increase in the angle of arrival of the wave, the angular error also increases.

2. With an increase in the angle of incidence of the wave at a constant porosity, the angular error also increases. In this case, the minimum error is observed at normal wave incidence (the angle of incidence is zero $\phi_0 = 0$), and the maximum is observed near the angles of incidence $\phi_0 = \pi/2$. There are ranges of wave incidence angles at which different porosities correspond to the same angular errors. For example, at $P = 10\%$, 50%, and the angle of incidence $\phi_0 = 0.8$ is rad, the angular error is 2.8 degrees. Moreover, at $\phi_0 < 0.8$ rad, the angular error at porosity $P = 10\%$ exceeds the angular error at $P = 50\%$, and $\phi_0 > 0.8$ vice versa at rad.

3. When the porosity changes from 0 to 100%, the angular error has three maxima and two minima at different wave incidence angles. The error maxima are concentrated in the porosity ranges (5%–15%), (45%–55%), and (95%–100%). In other areas of porosity values, we have minimal angular errors (figure 8).

4. The current state of the theory of aerodynamic heating of a dielectric [5, 8, 10, 14, 19, 20, 22, 23, and 25] does not allow establishing clear analytical relationships between the electrical characteristics of a dielectric and such external influencing factors as temperature, pressure, substance density, frequency of an external electromagnetic field. There are no prerequisites that make it possible to analytically formalize the temperature dependences of ε and $tg\delta$ of dielectrics with a partial change in the phase and state of the material as a result of high-temperature heating. On the way to solving this complex problem, there are significant theoretical difficulties that have not yet been overcome. The available information on the temperature dependences of ε and $tg\delta$ corresponds only to slow temperature changes and does not in any way reflect the properties of the material during thermal heating (pulsed mode of turning the laser beam on and off) or thermal shock.

Operating experience and numerous theoretical and experimental studies [1–4, 7, 8, 11, 13, 15, 16, 21, 24–26] have shown that when heated, fairings made of even the best dielectric materials noticeably change the radiation characteristics of the antennas they protect. In this case, the temperature dependences $\varepsilon(T)$ and $tan\delta(T)$ are determined experimentally for different thicknesses x of the dielectric layer. The obtained results are averaged over the thickness of the fairing wall. This method is time-consuming and inaccurate due to the low reliability of the dielectric parameters calculation and the disregard for heat distribution processes.

The approach developed in this work is based on the use of the law of thermal conductivity [6] with the distribution of temperature over the thickness of the dielectric layer (2) and the use of the known temperature dependence of the permittivity (1). The numerical coefficient in (1) is determined by the type of dielectric. In the general case, by choosing the coefficients A, A_1 in formula (6), one can write an expression for the permittivity of various ceramic materials.

The obtained relations, in contrast to the known solutions [5, 8, 20, 22, 23, 25] make it possible to represent the dielectric layer during aerodynamic heating in the form of a four-terminal network with elements of the resistance matrix (41)–(44), which makes it possible to increase the accuracy of calculations of antenna-radome systems for by taking into account the process of heat distribution over the thickness of the dielectric wall.

6. Conclusions

Development task solved models of a flat dielectric layer of a fairing under aerodynamic heating.

The scientific novelty lies in the fact that for the first time, a model of a flat dielectric layer in the form of a four-terminal network has been developed, which takes into account the gradient distribution of heat over the layer thickness. In this case, the four-terminal network is an inhomogeneous transmission line with an exponential change in the wave resistance along the thickness and depends on the temperature difference at the boundaries of the dielectric layer.

The practical value of the obtained results consists in increasing the accuracy of calculations of the radome during aerodynamic heating, which makes it possible to design antenna-radome systems for moving objects with the required radiation pattern under conditions of uneven heating of the radome wall. The research results can be used in the development of space communication and navigation systems with increased accuracy characteristics compared to existing analogs.

Prospects for further research are related to the development of analog and digital radome-antenna systems with increased angular accuracy, operating under conditions of aerodynamic heating.

References

[1] Akan V., Yazgan E.: Antennas for Space Applications: A Review. Advanced Radio Frequency Antennas for Modern Communication and Medical Systems, IntechOpen, 2020, [http://doi.org/10.5772/intechopen.93116].
[2] Chahat N.: A mighty antenna from a tiny CubeSat grows. IEEE Spectrum 55, 2018, 33–37 [http://doi.org/10.1109/MSPEC.2018.8278134].
[3] Deng J., Zhou G., Qiao Y.: Multidisciplinary design optimization of sandwich-structured radomes. Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 233(1), 2019, 179–189 [http://doi.org/10.1177/0954406218757268].
[4] Dippong T. et al.: Thermal behavior of Ni, Co and Fe succinates embedded in silica matrix. J. Therm. Analysis. Calorim. 136, 2019, 1587–1596 [http://doi.org/10.1007/s10973-019-08117-8].
[5] Escalera A. S. et al.: Effects of Radome Design on Antenna Performance in Transonic Flight Conditions. AIAA 2020-2187. AIAA Scitech 2020 Forum, 2020 [http://doi.org/10.2514/6.2020-2187].
[6] Gilchuk A. V., Khalatov A. A.: Theory of thermal conductivity. NTUU KPI named after Igor Sikorsky, 2017.
[7] Grinevich A. V., Lavrov A. V.: Evaluation of the ballistic characteristics of ceramic materials. Proceedings of VIAM 3(63), 2018, 95–102 [http://doi.org/10.18577/2307-6046-2018-0-3-95-102].
[8] Gyulmagomedov N. K.: Influence of the radiotransparent radome on characteristics of radar station. AIP Conference Proceedings 2318, 2021, 180001 [http://doi.org/10.1063/5.0036566].
[9] Korn G.: Handbook of mathematics for scientists and engineers: Definitions, theorems, formulas. Book on Demand, 2014.
[10] Li H. Y. et al.: Ameliorated Mechanical and Dielectric Properties of Heat-Resistant Radome Cyanate Composites. Molecules 25, 2020, 3117.
[11] Li H. Y. et al: Ameliorated Mechanical and Dielectric Properties of Heat-Resistant Radome Cyanate Composites. Molecules 25(14), 2020, 3117 [http://doi.org/10.3390/molecules25143117].
[12] Lu Y. et al.: A Study on the Electromagnetic–Thermal Coupling Effect of CrossSlot Frequency Selective Surface. Materials 15, 2022, 640 [http://doi.org/10.3390/ma15020640].
[13] Meyer G. J.: Polyurethane Foam: Dielectric Materials for Use in Radomes and Other Applications. General Plastics Manufacturing Company, 2015.
[14] Nair R. U. et al.: Temperature-dependent electromagnetic performance predictions of a hypersonic streamlined radome. Prog. electromagn. Res. 154, 2015, 65–78.
[15] Narendara S., Gopikrishna R.: Evaluation of structural integrity of tactical missile ceramic radomes under combined thermal and structural loads. Procedia Structural Integrity 14, 2019, 89–95.
[16] NASA Outgassing Data for Selecting Spacecraft Materials, https://outgassing.nasa.gov (available: April 20, 2020).

[17] Öziş E. et al.: Metamaterials for Microwave Radomes and the Concept of a Metaradome: Review of the Literature. International Journal of Antennas and Propagation 2017, ID1356108 [http://doi.org/10.1155/2017/1356108].
[18] Plonus M.: Electronics and Communications for Scientists and Engineers, 2020 [http://doi.org/10.1016/C2018-0-00442-9].
[19] Raveendranath U. N. et al.: Temperature-Dependent Electromagnetic Performance Predictions of a Hypersonic Streamlined Radome. Progress In Electromagnetics Research 154, 2015, 65–78.
[20] Romashin A. G. et al.: Radiotransparent fairings for aircraft. National Aerospace University, Kharkov 2003.
[21] Seckin S. et al.: Dielectric Properties of Low-Loss Polymers for mmW and THz Applications. International Journal of Infrared and Millimeter Waves 40, 2019, 557–573 [http://doi.org/10.1007/s10762-019-00584-2].
[22] Tahseen H. U. et al.: Design of FSS-antenna-radome system for airborne and ground applications. LET Communications, 2021 [http://doi.org/10.1049/cmu2.12181].
[23] Tahseen H. U. et al.: Design of FSS-antenna-radome system for airborne and ground applications. IET Commun. 2021, 15, 1691–1699, [http://doi.org/10.1049/cmu2.12181].
[24] Xu W. et al.: Study on the electromagnetic performance of inhomogeneous radomes for airborne applications part 1: Characteristics of phase distortion and boresight error. IEEE Transactions on Antennas and Propagation 65(6), 2017, 3162–3174.
[25] Ya M. et al.: Physics of heating microwave dielectrics of aircraft and their protection. SSGA, Novosibirsk 2008.
[26] Zhang H. X. et al.: Massively Parallel Electromagnetic–Thermal Cosimulation of Large Antenna Arrays. IEEE Antennas Wire. Propag. Lett. 19, 2020, 1551–1555.

D.Sc. Valerii Kozlovskiy
e-mail: valerey@ukr.net

Research interests: radio electronics, ultra-high frequencies techniques, nonregular distributed circles.

http://orcid.org/0000-0003-0234-415X

D.Sc. Valeriy Kozlovskiy
e-mail: vvkzeos@gmail.com

Research interests: telecommunication systems, programmable logic controllers, research network security, features of the communication lines organization and operation.

http://orcid.org/0000-0002-8301-5501

Oleksii Nimych
e-mail: aleksei.nimich@gmail.com

Research interests: cybersecurity, software engineering, antenna systems

http://orcid.org/0000-0003-1759-7088

Liudmyla Klobukova
e-mail: klp@nau.edu.ua

Research interests: telecommunications and information technologies, multichannel communication systems, orthogonal modulation

http://orcid.org/0000-0001-9799-4387

Ph.D. Natalia Yakymchuk
e-mail: n.yakymchuk@lntu.edu.ua

Research interests: diagnostics and control of the telecommunication networks state, end-to-end diagnostics, congestion management.

http://orcid.org/0000-0002-8173-449X

http://doi.org/10.35784/iapgos.5355 *received: 24.09.2023 | revised: 11.11.2023 | accepted: 01.12.2023 | available online: 20.12.2023*

MICROWAVE MIXER ON RECTANGULAR WAVEGUIDES PARTIALLY FILLED BY DIELECTRIC

Vitaly Pochernyaev[1], Nataliia Syvkova[1], Mariia Mahomedova[2]
[1]National Academy of Security Service of Ukraine, Kyiv, Ukraine, [2]Kyiv Professional College of Communications, Kyiv, Ukraine

Abstract. *The article investigates and calculates the characteristics of microwave mixers on rectangular waveguides partially filled by dielectric. Presents diagrams of promising combined microwave radio engineering systems - two options for constructing mobile digital troposcatter-radiorelay stations, the antenna-feeder paths of which are implemented on rectangular waveguides partially filled by dielectric. At research of microwave mixers, suppression and the use of the mirror frequency are taken into account. The analysis of researches of microwave mixers is carried out. The design of a balanced-type microwave mixer based on rectangular waveguides partially filled by dielectric is developed. The mixer is used to convertion the microwave signal into an intermediate frequency signal. The signal conversion of the mixers takes place on the non-linear active resistance of the semiconductor diode. In article, an open nonlinear structure is used as such a diode. The following main parameters of microwave mixers are investigated: conversion losses, noise factor, operating frequency band, signal suppression at the mirror frequency. The conversion losses are determined for various mirror channel suppression conditions. Phase methods of mirror frequency suppression are considered, which are most suitable for the waveguide implementation of microwave mixers. A scheme of a microwave mixer of a balance type with a phase method for suppressing the mirror frequency is presented. The article notes that for significant suppression of the mirror frequency of more than 30 dB, a double frequency conversion mixer is used. A diagram of a slotted bridge based on rectangular waveguides partially filled by dielectric is presented. The dependences of the input impedance of the mixer, the impedance of the mixer at the intermediate frequency, the impedance of the mixer at the mirror frequency by the power of the local heterodyne are plotted.*

Keywords: microwave mixer, mobile digital troposcatter-radiorelay station, rectangular waveguide, slot bridge, conversion loss, mirror frequency

MIKSER MIKROFALOWY NA PROSTOKĄTNYCH FALOWODACH CZĘŚCIOWO WYPEŁNIONYCH DIELEKTRYKIEM

Streszczenie. *W artykule zbadano i obliczono charakterystyki mieszaczy mikrofalowych na falowodach prostokątnych częściowo wypełnionych dielektrykiem. Przedstawiono schematy obiecujących połączonych mikrofalowych systemów inżynierii radiowej – dwóch opcji budowy mobilnych cyfrowych troposferycznych stacji radiowych, których tory antenowo-zasilające są realizowane na prostokątnych falowodach częściowo wypełnionych dielektrykiem. W badaniach mieszaczy mikrofalowych uwzględniono tłumienie i wykorzystanie częstotliwości lustrzanej. Przeprowadzono analizę badań mieszaczy mikrofalowych. Opracowano projekt mieszacza mikrofalowego typu zrównoważonego opartego na prostokątnych falowodach częściowo wypełnionych dielektrykiem. Mieszacz służy do konwersji sygnału mikrofalowego na sygnał o częstotliwości pośredniej. Konwersja sygnału mieszacza odbywa się na nieliniowej rezystancji czynnej diody półprzewodnikowej. W artykule jako taką diodę zastosowano otwartą strukturę nieliniową. Badane są następujące główne parametry mieszaczy mikrofalowych: straty konwersji, współczynnik szumów, pasmo częstotliwości roboczej, tłumienie sygnału na częstotliwości lustrzanej. Straty konwersji są określane dla różnych warunków tłumienia kanału lustrzanego. Rozważono fazowe metody tłumienia częstotliwości lustrzanej, które są najbardziej odpowiednie do falowodowej implementacji mieszaczy mikrofalowych. Przedstawiono schemat mieszacza mikrofalowego typu balansowego z fazową metodą tłumienia częstotliwości lustrzanej. W artykule zauważono, że w celu znacznego tłumienia częstotliwości lustrzanej o ponad 30 dB stosuje się mieszacz z podwójną konwersją częstotliwości. Przedstawiono schemat mostka szczelinowego opartego na prostokątnych falowodach częściowo wypełnionych dielektrykiem. Wykreślono zależności impedancji wejściowej mieszacza, impedancji mieszacza dla częstotliwości pośredniej, impedancji mieszacza dla częstotliwości lustrzanej od mocy lokalnej heterodyny.*

Słowa kluczowe: mieszacz mikrofalowy, mobilna cyfrowa troposferyczna stacja radiowa, falowód prostokątny, mostek szczelinowy, straty konwersji, częstotliwość lustrzana

Introduction

In areas of natural disasters and man-made disasters, in the event of emergency and emergency situations, as well as in the conditions of hostilities or special ground operations, the most likely option for organizing a terrestrial communication line in the microwave range may be to build a combination of troposcatter and radiorelay communication stations. At present, despite the widespread use of satellite communication systems, troposcatter stations are used both in special-purpose networks and in departmental and corporative networks. Higher survivability in a difficult jamming environment is an advantage of troposcatter communications over other microwave radio systems, including satellite ones [1–3, 6, 7, 17–20].

Schemes of mobile digital troposcatter-radiorelay stations (MDTRRS), which are combined microwave radio systems, are shown in Fig. 1 and Fig. 2 [12, 13]. Such a combined microwave radio system provides an over-the-horizon communication mode in the presence of a troposcatter component and a line-of-sight mode in the presence of a radiorelay component, which makes it possible to build a transport network with different distances between control and relay points.

On Fig. 1 shows the microwave transceiver path of a MDTRRS (option 1), where: *I* and *II* – antennas of troposcatter component; *A* and *B* – antennas of radiorelay component; *V* and *H* – vertical and horizontal polarization; *PS 1TC* and *PS 2TC* – polarization selectors of troposcatter component; *D1* and *D2* – duplexers; *DP* – distributor of power; *PD1* and *PD2* – power dividers; *FBFET 1*, *FBFET 2*, *FBFET 3* and *FBFET 4*, – frequency bandpass filters with electronically tunable; *MR 1TC*, *MR 2TC*, *MR 3TC* and *MR 4TC* – microwave receivers of troposcatter component; *MM 1TC*, *MM 2TC*, *MM 3TC* and *MM 4TC* – microwave mixers of troposcatter component; *PS 1RC* and *PS 2RC* – polarization selectors of radiorelay component; EH - exciter-heterodyne; *OBFFT 1RC* and *OBFFT 2RC* – output bandpass filters with frequency tunable of radiorelay component; *MR 1RC*, *MR 2RC* – microwave receivers of radiorelay component; *MM 1RC* and *MM 2RC* – microwave mixers of radiorelay component; *MT1* and *MT2* – microwave transmitters; *CP*– control panel; *DSTC* – device of space-time coding; *PIFA1*, *PIFA2*, *PIFA3* and *PIFA4* – previous intermediate frequency amplifier; *M* – modem; *M-D RC* – multiplexer-demultiplexer of radiorelay component, *S1*, *S2* – switch, *SCP1*, *SCP2* – switch control panel.

Figure 2 shows the microwave transceiver path of a MDTRRS (option 2), where: *I* and *II* – antennas of troposcatter component; *A* and *B* – antennas of radiorelay component; *V* and *H* – vertical and horizontal polarization; *PS 1TC* and *PS 2TC* – polarization selectors of troposcatter component; *D1* and *D2* – duplexers; *DP* – distributor of power; *BFFT 1*, *BFFT 2*, *BFFT 3* and *BFFT 4* – bandpass filters with frequency tunable; *MR 1TC*, *MR 2TC*, *MR 3TC and MR 4TC* – microwave receivers of troposcatter component; *MM 1TC*, *MM 2TC*, *MM 3TC* and *MM 4TC* – microwave mixers of troposcatter component; *PS 1RC* and *PS 2RC* – polarization selectors of radiorelay component; *OBFFT 1RC* and *OBFFT 2RC* – output bandpass filters with frequency tunable of radiorelay component; *RT 1RC* and *RT 2RC* – receiver-transmitter of radiorelay component; *EA*

1TC and *EA 2TC* – equivalents antennas of troposcatter component; *MT 1TC* and *MT 2TC* – microwave transmitters of troposcatter component; EH – exciter-heterodyne; *PIFA1*, *PIFA2*, *PIFA3* and *PIFA4* – previous intermediate frequency amplifier; *M-D RC* – multiplexer-demultiplexer of radiorelay component; *M* – modem; *AC* – automatic control; *PAS* – power adaptation system; *FAS* – frequency adaptation system; *DSTC* – device of space-time coding.

To reduce the weight and size parameters and increase the electrical strength, the microwave paths of such stations can be implemented on rectangular waveguide partially filled by dielectric (RWPFD). However, a complete range of active devices based on RWPFD has not been developed, including microwave mixers. The schemes on Fig.1, 2 contain a six microwave mixers.

The aim of the article is to research the characteristics and parameters of microwave mixers on RWPFD.

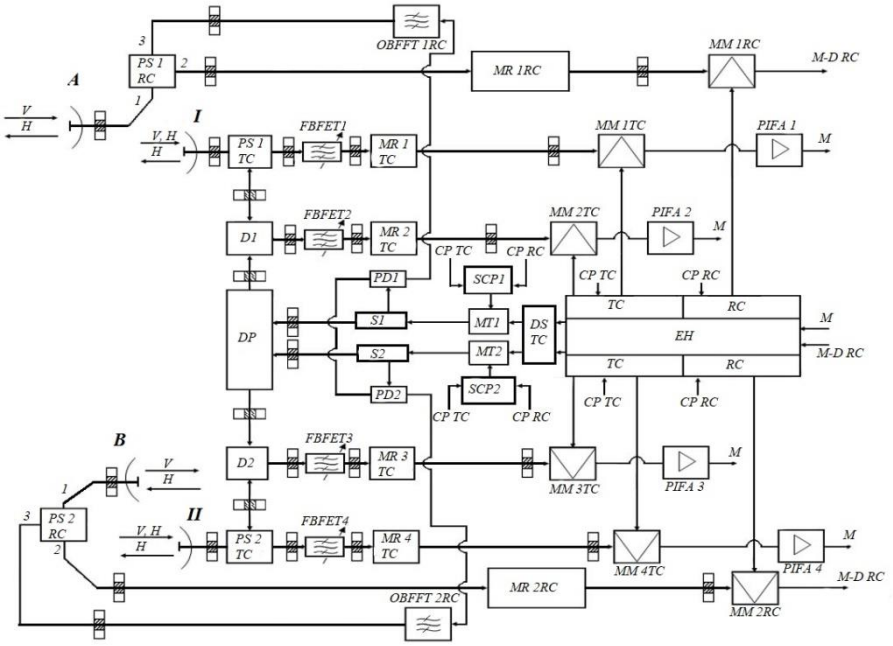


Fig. 1. Transmitting and receiving microwave paths MDTRRS (option 1)

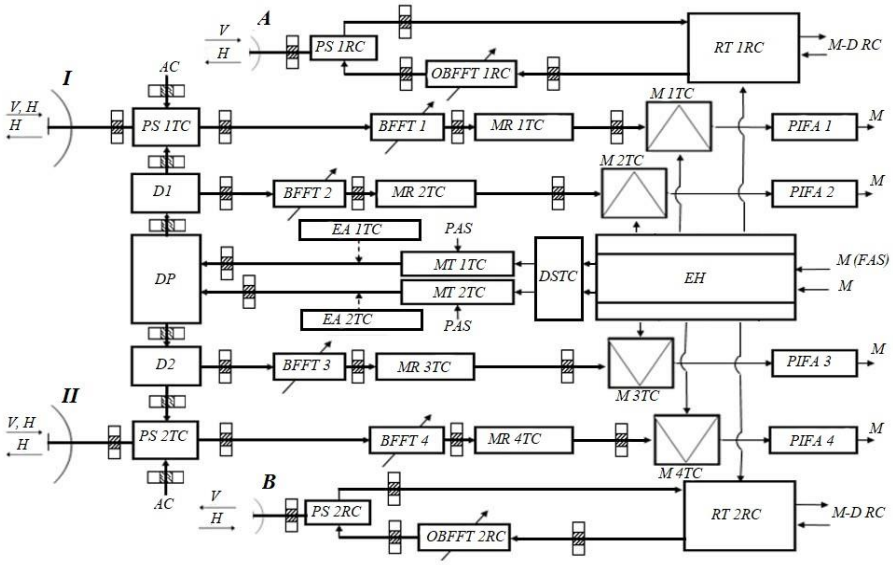


Fig. 2. Transmitting and receiving microwave paths MDTRRS (option 2)

1. Analysis of microwave mixer research

The performed analysis of mixers shows that studies will be introduced in low noise directions for software defined radios [15, 16]. The most promising "technique" in the suppression of the mirror frequency is the phase method [5, 9, 11]. The technique for measuring the characteristics and parameters of mixer models, microwave and extremely high frequencies ranges is based on the use of a digital signal generator and a spectrum analyzer described in [4]. The results of experimental studies of promising designs of balanced microwave mixers are given in [8].

2. Main part

A microwave mixer is needed to convert the microwave signal into an intermediate frequency (IF) signal. The conversion of the mixer signal occurs on the active non-linear resistance of the semiconductor diode. In this article, an open non-linear structure (ONS) is used as a semiconductor diode, placed in a dielectric plate RWPFD. The following main characteristics and parameters of the microwave mixer were studied: conversion loss, noise factor, operating frequency band, suppression of the mirror channel. This study does not include the study of the combinational components suppression, but only considers the suppression and the use of the mirror frequency (mirror channel).

On Fig. 3 shows the design of the microwave mixer at RWPFD: *a*) - general view; *b*) - face view; *c*) top view.

The design of the balanced mixer at RWPFD is shown in Fig. 3.

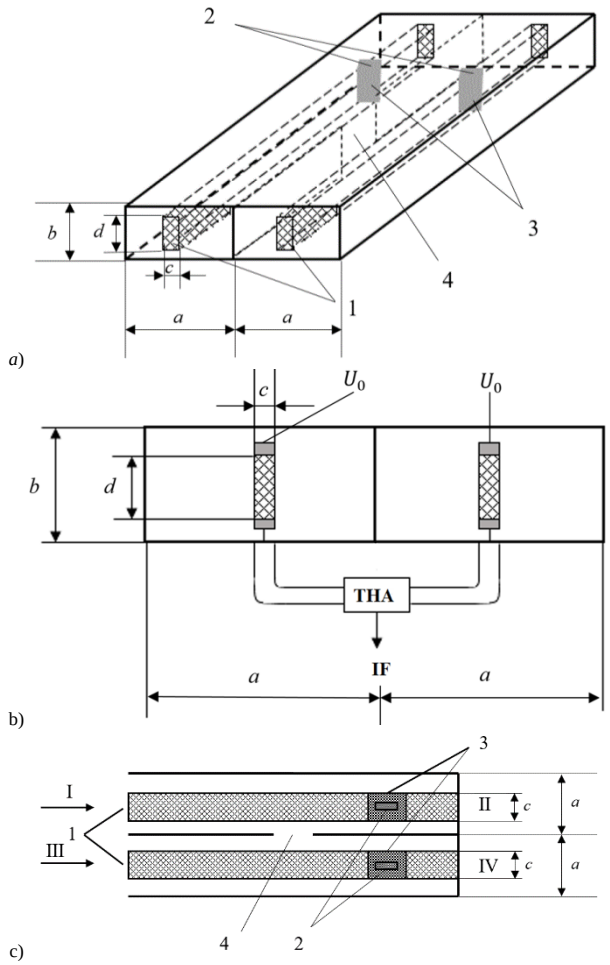


Fig. 3. Slotted bridge on RWPFD: 1– dielectric plates; 2 – non-radiating slots; 3 – open nonlinear structures; 4 – connection hole

Inputs I and III are supplied with a useful signal and a local heterodyne signal, respectively. Outputs II and IV are shortened, the control voltage U_0 is supplied through the non-radiating slots 2. The connection hole 4 provides the local heterodyne signal with an additional phase shift $\pi/2$. In the second channel, the additional phase shift $\pi/2$ through the connection hole 4 acquires a useful signal. Therefore, the instantaneous phases of the IF currents on the first and second ONS are equal:

$$\varphi_1 = \left(\varphi_h - \frac{\pi}{2}\right) - \varphi_s,\ \varphi_2 = \varphi_h - \left(\varphi_s - \frac{\pi}{2}\right)$$

From this it can be seen that the phases φ_1 and φ_2 differ by π, therefore, when the ONS is turned on in the opposite direction, we get the addition of currents at the input of the IF amplifier. Note that the noise oscillations turn out to be in-phase and, when the ONS are connected back to back, are mutually compensated. Balanced circuit makes it easier to tune the mixer and increases the decoupling of the signal and heterodyne frequencies.

On Fig. 4 shows a diagram of a transformer hybrid adder (THA) used as an adder and a matching device.

The circuit in Fig. 4 makes it possible to implement a wide range of matched resistances *r* with relatively small and quite acceptable values of *n* and *k* from the point of view of obtaining wide bandwidths. This scheme, when using two identical transformers with those indicated in Fig. 4 transformation ratios works in a symmetrical version at $n = 1 \pm \sqrt{2}$.

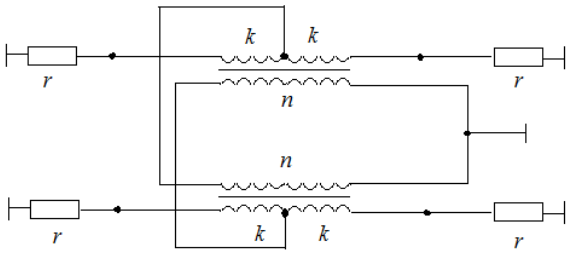


Fig. 4. Scheme of THA

The frequency range of the developed slotted bridge on RWPFD, as well as the microwave mixer, is determined by:

$$\frac{2\Delta\lambda}{\lambda_0} \approx \left(\frac{2\Delta\theta}{\theta_0}\right)\left[\left(\varepsilon_{eff1} - \left(\frac{\lambda_0}{2a}\right)^2\right)\left(\varepsilon_{eff2} - \left(\frac{\lambda_0}{4a}\right)^2\right)\right]^{\frac{1}{2}} \quad (1)$$

$$\varepsilon_{eff1} = 1 + (\varepsilon_r - 1)\frac{cd}{ab}\left[1 + \frac{\sin(\pi c/a)}{\pi c/a}\right]$$

$$\varepsilon_{eff2} = 1 + (\varepsilon_r - 1)\frac{cd}{ab}\left[1 - \frac{\sin(2\pi c/a)}{2\pi c/a}\right]$$

where λ_0 – frequency band center wavelength; $2\Delta\theta/\theta_0$ – value specifying the level of crosstalk to be set; ε_{eff1}, ε_{eff2} – effective permittivity of quasi-H_{10} and quasi-H_{20} waves, respectively; *a*, *b* – dimensions of a rectangular waveguide; *c*, *d* – dimensions of the dielectric plate; ε_r – relative permittivity of a dielectric plate.

The values $2\Delta\theta/\theta_0$ can be found using the following expressions:

$$\Delta\theta = \arcsin 10^{-0.05(\alpha-\Delta\alpha)} - \arcsin 10^{-0.05(\alpha+\Delta\alpha)}$$

$$\theta_0 = (2n-1)\pi/2,\ n = 1, 2, 3, \ldots$$

A standard waveguide slot bridge (WSB) on empty rectangular waveguides has a bandwidth of 9.2%. The throughput capacity of the developed WSB on RWPFD with dimensions *c*/*a* = 0.12 and *d*/*b* = 0.8 according to formula (1) is as follows:

– for $\varepsilon_r = 2.4$ we have $2\Delta\lambda/\lambda_0 = 10.8\%$

– for $\varepsilon_r = 4$ we have $2\Delta\lambda/\lambda_0 = 12.4\%$

– for $\varepsilon_r = 9.6$ we have $2\Delta\lambda/\lambda_0 = 17.1\%$

Let's move on to the analysis of other main parameters of the mixer. The mixer conversion loss L_s is defined as the ratio of the input power of the microwave signal $P_{in.s}$ to the output power of the signal at the intermediate frequency P_{IF}. Losses are usually expressed in decibels:

$$L_s = 10\lg(P_{in.s}/P_{IF})$$

The mixer conversion losses consist of the ONS conversion losses L_d, the losses of the ONS inconsistency at the input and output, as well as losses in the passive elements of the mixer, for example, a slotted bridge.

The conversion loss L_d is related to the properties of the semiconductor material in such a way:

$$L_d \sim \sqrt{\varepsilon_d}$$

where ε_d – relative dielectric permittivity ONS.

Noise factor:

$$K_N = 10\lg\frac{P_{in.s}/P_{in.N}}{P_{out.s}/P_{out.N}}$$

where $P_{in.s}$ and $P_{out.s}$ – nominal input and output signal power; $P_{in.N}$ and $P_{out.N}$ – nominal input and output noise power.

The noise factor of the mixer is a generalized parameter that takes into account the noise factor of the ONS $K_{N.d}$ and conversion losses, as well as the noise factor of the intermediate frequency amplifier (IFA) $K_{N.IFA}$. The noise factor of a mixer with IFA is described by the equation: $K_N = L_S(K_{N.d} + K_{N.IFA} - 1)$.

Usually $K_{N.IFA}$=1.4 (or 1.5 dB).

Suppression of the mirror channel is determined by the ratio of the signal power of the IF to the signal power on the mirror channel P_m and is expressed in decibels:

$$K_{p.m} = 10\lg(P_{IF} / P_m)$$

The conversion losses can be determined using the theory of linear electrical circuits. In the simplest case, it is assumed that the following changes in nonlinear conductivity occur in the ONS under the influence of the local heterodyne voltage:

$$g = g_0 + 2\sum_{n=1}^{\infty} g_n \cos n\omega t \qquad (2)$$

where ω – oscillation frequency of the local heterodyne; g_0 – the constant component of the conductivity of the ONS; g_n – fourier component of conductivity at frequencies $n\omega$.

To determine the transformation losses, introduce the notation:

$$\gamma_n = g_n / g_0$$

Determine the conversion loss for different conditions for suppressing the mirror channel:

– the mirror channel is matched to the load:

$$L_m = 2\left(1+\sqrt{1-\eta_c}\right)/\left(1-\sqrt{1-\eta_c}\right), \eta_c = 2\gamma_1^2/(1+\gamma_2)$$

– the mirror channel on a short-circuited load:

$$L_m = \left(1+\sqrt{1-\eta_k}\right)/\left(1-\sqrt{1-\eta_k}\right), \eta_k = \gamma_1^2$$

– the mirror channel on open load:

$$L_m = \left(1+\sqrt{1-\eta_p}\right)/\left(1-\sqrt{1-\eta_p}\right),\ \eta_p = \gamma_1^2(1-\gamma_2)/(1-\gamma_1^2)(1+\gamma_2)$$

Determine the input conductivity depending on the load of the mirror channel:

– the mirror channel is matched to the load:

$$g_{in} = g_0\sqrt{(1-\gamma_2)(1+\gamma_2-2\gamma_1^2)}$$

– the mirror channel on a short-circuited load:

$$g_{in} = g_0\sqrt{(1-\gamma_1^2)}$$

– the mirror channel on open load:

$$g_{in} = g_0\sqrt{(1-\gamma_2^2)}\sqrt{(1-\gamma_2)(1+\gamma_2-2\gamma_1^2)/(1-\gamma_1^2)}$$

Determine the IF conductivity:

– the mirror channel matched to the load:

$$g_{our} = g_0\sqrt{(1+\gamma_2-2\gamma_1^2)/(1+\gamma_2)}$$

– the mirror channel on a short-circuited load:

$$g_{our} = g_0\sqrt{(1-\gamma_1^2)}$$

– the mirror channel on open load:

$$g_{our} = g_0\sqrt{(1-\gamma_2)^2(1+\gamma_2-2\gamma_1^2)(1-\gamma_1^2)}$$

If the indicated values are calculated with a short-circuited load through the mirror channel, i.e. $Y_m = \infty$, then the impedance of the mixer IF is lower resistance than with a matched load and an open channel of the mirror frequency. This does not require agreement with the previous IFA (PIFA). Therefore, the noise factor of the mixer with PIFA is lower, although theoretically the mixer conversion loss with an open channel of the mirror frequency is less than with a short-circuited one. In this case, the ONS is a symmetrical four-terminal network and its initial conductivity is equal to the conductivity at the IF.

Taking into account only the frequencies (signal, intermediate, mirror) and conductivity according to formula (2), we obtain the complex conductivity matrix of the ONS:

$$\begin{bmatrix} g_0 + ja_0\omega_s & g_1 + ja_1\omega_s & g_2 + ja_2\omega_s \\ g_{-1} + ja_{-1}\omega_{IF} & g_0 + ja_0\omega_{IF} & g_1 + ja_1\omega_{IF} \\ g_{-2} + ja_{-2}\omega_m & g_{-1} + ja_{-1}\omega_m & g_0 + ja_0\omega_m \end{bmatrix} \qquad (3)$$

where $\omega_s = \omega_{IF} + \omega_h$; $\omega_m = \omega_h + \omega_{IF}$.

The conductivities included in matrix (3) are calculated according to the method of [10]. Note that two waves quasi-H_{10} and quasi-H_{20} propagate in the coupling region of the slot bridge. In the calculation of reactive conductivities, the main contribution to the formation of the local field is made by the quasi-H_{30} wave.

The transverse electrical eigenfunctions of these waves, through which the conductivities of matrix (3) are determined written as follows:

$$\bar{\mathcal{E}}_{h_{10}} = \left(\sqrt{128/ab\left(64+q^2+p^2+q^2p^2\right)}/\chi_{h_{10}}\right)*\mathcal{F}$$

$$\mathcal{F} = \left\{\left[\left(\frac{\pi}{a}\right)\sin\frac{\pi x}{a} - \left(\frac{p\pi}{2a}\right)\sin\frac{\pi x}{a}\cos\frac{2\pi y}{b} - \left(\frac{3q\pi}{ba}\right)* \right.\right.$$
$$*\sin\frac{3\pi x}{a}\sin\frac{\pi x}{a}\cos\frac{2\pi y}{b} - \left(\frac{3q\pi}{ba}\right)\sin\frac{3\pi x}{a} +$$
$$\left.+\left(\frac{3qp\pi}{16a}\right)\sin\frac{3\pi x}{a}\cos\frac{2\pi y}{b}\right]\bar{y}^0 + \left[\left(\frac{p\pi}{b}\right)\cos\frac{\pi x}{a}* \right.$$
$$\left.\left.*\sin\frac{2\pi y}{b} - \left(\frac{2qp\pi}{8b}\right)\cos\frac{3\pi x}{a}\sin\frac{2\pi y}{b}\right]\bar{x}^0\right\}$$

$$\bar{\mathcal{E}}_{h_{20}} = \left(\sqrt{128/ab\left(64+q^2+p^2+q^2p^2\right)}/\chi_{h_{20}}\right)*$$
$$*\left[\left(p\tau\cos 2\delta x\sin 2\tau y - \left(\frac{qp\tau}{8}\right)Q_1\sin 2\tau y\right)\bar{x}_0 + (2\delta\sin 2\delta x - \right.$$
$$\left.-p\delta\sin 2\delta x\cos 2\tau y - \left(\frac{q\delta}{3}\right)\sin 4\delta x\cos 2\tau y)\bar{y}_0\right]$$
$$Q_1 = \frac{2}{3}\cos 4\delta x - 2$$

$$\bar{\mathcal{E}}_{h30} = \sqrt{128/ab\left(64+q^2+p^2+q^2p^2\right)}*\frac{1}{k_{h30}}*\mathcal{J}$$

$$\mathcal{J} = \left\{\left[\frac{p\pi}{b}\cos\frac{3\pi x}{a}\sin\frac{2\pi y}{b} - \frac{qp\pi}{16b}\left(\cos\frac{5\pi x}{a} - 2\cos\frac{\pi x}{a}\right)* \right.\right.$$
$$\left.*\sin\frac{2\pi y}{b}\right]\bar{x}^0 + \left[\frac{3\pi}{a}\sin\frac{3\pi x}{a} - \frac{3p\pi}{2a}\sin\frac{3\pi x}{a}\cos\frac{2\pi y}{b} - \right.$$
$$-\frac{q\pi}{8a}\left(\frac{5}{2}\sin\frac{5\pi x}{a} - \sin\frac{\pi x}{a}\right) +$$
$$\left.\left.+\frac{qp\pi}{16a}\left(\frac{5}{2}\sin\frac{5\pi x}{a} - \sin\frac{\pi x}{a}\right)\cos\frac{2\pi y}{b}\right]\bar{y}^0\right\}$$

where $\chi_{h_{10}}$, $\chi_{h_{20}}$, $\chi_{h_{30}}$ – transverse wavenumbers of waves quasi-H_{10}, quasi-H_{20}, quasi-H_{30} [20].

Consider phase methods for suppressing mirror frequencies.

So, for example, four branches of the receiver of the tropo-scatter component of the MDTRRS contain four microwave mixers (Fig. 1, 2). In addition to the main conversion in microwave mixers, there is a parasitic conversion to the frequency of the mirror channel.

The appearance of the mirror channel occurs due to the transmission of the signal range from one radio frequency region to another. In the scheme of the balance-type mixer in Fig. 5, a microwave mixer is used, which is shown in Fig. 3. The circuit in Fig. 5 should be highly symmetrical. In this case, the suppression of the mirror channel is carried out by the phase method.

If the phase of the converted signal is shifted by less than 5 degrees, and when the signal amplitude difference is 1 dB from two non-linear elements, then the microwave mixer circuit can suppress the mirror frequency signal by 25 dB.

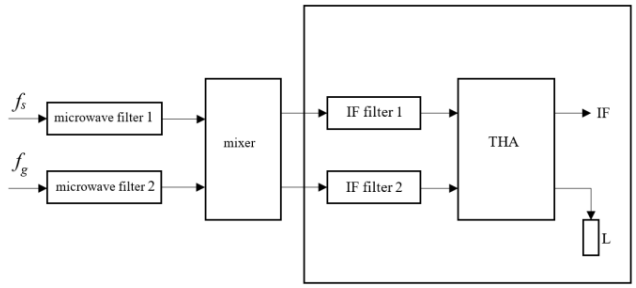


Fig. 5. Scheme of a microwave mixer of a balance type with a phase method of suppression

The scheme in Fig. 5 includes: IF filter 1, IF filter 2 – intermediate frequency filters; L – load; IF – intermediate frequency; THA – transformer hybrid adder. The mixer is built according to the design of Fig. 3.

The loss L_m conversion of mirror frequency is a function of factor K_m conversion the signal to mirror frequency:

$$K_m = (P_m / P_s)\cos\left[2\pi(2f_s - f_{sf})t + (2\varphi_s - \varphi_{sf})\right]$$

where P_m – power of mirror frequency; P_s – input signal power; f_s – input signal frequency; f_{sf} – frequency synthesizer frequency; φ_s – input signal phase; φ_{sf} – frequency synthesizer signal phase.

It should be taken into account that the conversion loss and therefore the efficiency of the mirror frequency signal suppression, depends on the characteristics of the non-linear elements.

On Fig. 6 shows a fragment of the IF included in the microwave mixer of the balanced type. The addition of devices for controlling and controlling the amplitude and phase (APCMD 1, APCMD 2) makes it possible to improve the suppression of the mirror channel.

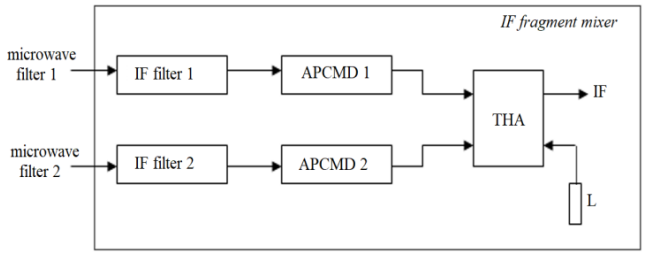


Fig. 6. Fragment of the IF included in the microwave mixer of the balance type with the phase method of suppression

These devices make it possible to provide a phase deviation of 1 degree from two non-linear elements and an amplitude of their difference of 0.1 dB. This allows an additional 15 dB of mirror frequency rejection to be added. Therefore, the total rejection of the mirror frequency signal using APCMD 1, APCMD 2 is approximately 40 dB.

The dependence of losses L_m and L_s in the frequency range is shown in Fig. 7.

Note that if significant mirror frequency rejection of more than 30 dB is required, a double frequency conversion mixer is used. Such a scheme is implemented by switching on two mixers in series: the first one transfers the signal to the high first IF, at which it is easy to suppress the mirror frequency with filters, and then the second mixer turns the high IF into a low IF, on which further signal processing takes place. Note that the frequency range in Fig. 7 corresponds to the frequency range of a mobile digital troposcatter station.

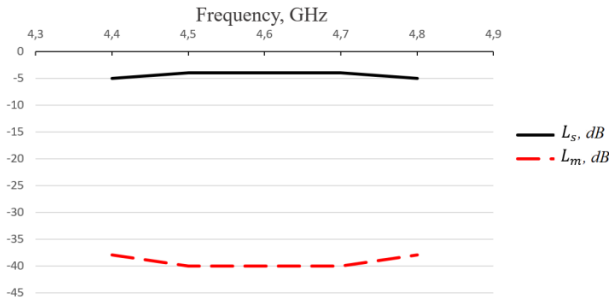


Fig. 7. Dependence of losses L_m and L_s in the frequency range

To obtain the minimum heterodyne power through the signal input and low mixer noise factor, a sufficiently high isolation between the heterodyne and signal input is required. For example, 10 dB isolation increases the noise factor of the mixer by about 10% and 10% of the heterodyne power is radiated. At 20 dB isolation, the noise factor degradation is 1% and about 1% of the local heterodyne power is radiated through the signal input of the mixer.

As can be seen in Fig. 8a, the dependence of the value of the input resistance R_{in} and the intermediate frequency R_{IF} on the power of the local heterodyne P_h in the matched load mode via the mirror channel is shown. On Fig. 8b shows the input resistance values R_{in} and IF R_{IF} depending on the power of the local heterodyne P_h in the short circuit mode ($Y_m = \infty$) on the mirror channel. On Fig. 8c shows the values of the input resistance R_{in} and IF R_{IF} from the power of the local heterodyne P_h in the idle mode ($Y_m = 0$) via the mirror channel. The conversion loss L_s does not depend on the external bias. In calculations, the power of the local heterodyne is related to the voltage U_h by the following expression: $2P_h = U_h^2 / R_{in}$.

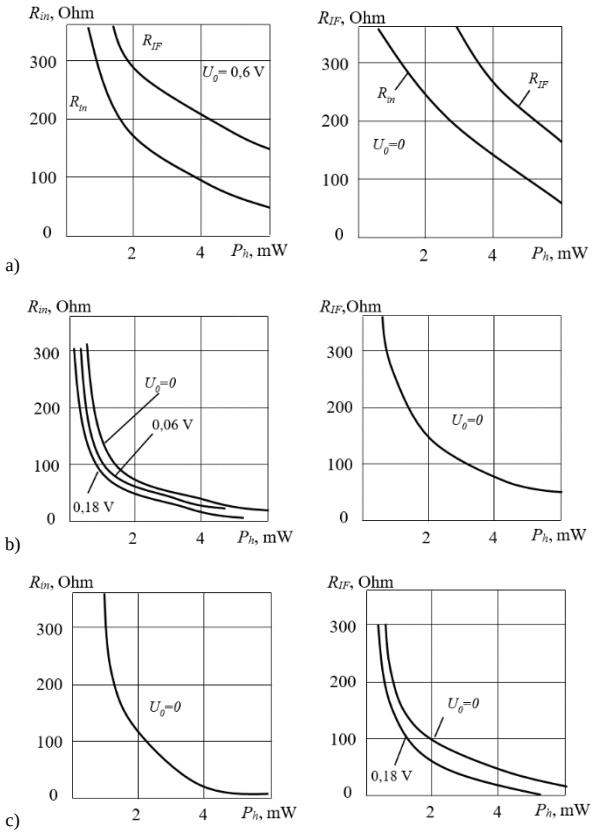


Fig. 8. The dependence of the input resistance and the IF resistance from a heterodyne power

It should be noted that the calculation and matching of the ONS should be carried out in microwave mixers, taking into account filters through the mirror channel, if any are provided in the circuits of microwave receivers. Otherwise, when the mixer is turned on in the receiving device, in which there is a filter, an inconsistency of the ONS will occur.

For example, the input conductivity of the ONS, provided that the junction capacitance C_j is a constant value, the following:

$$Y_{in} = \frac{1}{R_{in}} + j\omega C_j \approx 8.7\times10^{-3} + j\times7\times10^{-3}$$

Then the local heterodyne power consumed by the ONS: $P_h = U_h^2 Y_{in} / 2 \approx 2.6$ mW. Therefore, it is necessary to coordinate the ONS with the transmission line according to the input value Y_{in}.

When designing a microwave mixer, it is necessary to take into account the way in which the mirror frequency is usefully used. Conversion and noise factor losses can be minimized by proper choice of reactive load at signal and mirror frequencies. However, most often this is very difficult to implement, especially if the mirror and signal frequencies are close. There are two methods for solving this problem: the introduction of frequency-selective circuits and the introduction of phase relationships between signals. The circuit assembled on the basis of the first method can operate in a narrow frequency band. In addition, if the difference between the mirror frequency and the signal frequency is small, then a very high quality filter with low losses is needed. Examples of the implementation of such schemes are known, which made it possible to obtain conversion losses of up to 3.5 dB.

It should be noted that there are two signals at the mirror frequency in the mixer: the incoming signal to the input of the mixer from the antenna and the signal formed in the mixer due to the transformation of the input signal. If an external signal with a frequency ω_m arrives at the ONS through the signal input, then, interacting with the oscillations of the local heterodyne, an intermediate frequency signal is formed $\omega_s' = \omega_m - \omega_h$, $\varphi_s' = \varphi_m - \varphi_h$. The phase of this IF φ_s' is not correlated with the phase of the useful signal φ_s, although it does not differ in frequency from the useful signal and is an interference that cannot be eliminated without the use of special measures.

3. Conclusions

In conclusion, we note that in the short circuit mode through the mirror channel, the input impedance decreases significantly and reaches 50 Ohm at a power of 3×10^{-3} W that is, two times less than in the matched load mode. If we take into account that the conversion loss in this case drops to 1.8 dB, then the short circuit mode in the mirror channel is the preferred mode of operation of a balanced microwave mixer. In idle mode, the mirror channel has an even lower conversion loss of 1.2 dB. It can be seen from the graphs that the resistance to the input R_{in} in the matched load mode is 50 Ohm at the local heterodyne power $P_h = 6\times10^{-3}$ W, and with external bias $U_0 = 0.18$ V the input resistance is 50 Ohm at the local heterodyne power $P_h = 4\times10^{-3}$ W.

References

[1] Bastos L., Wietgrefe H.: Geographical Analysis of Highly Deployable Troposcatter Systems. Performance IEEE Military communications conference – MILCOM, San Diego, 2013, 661–667.

[2] Bastos L., Wietgrefe H.: Highly-deployable troposcatter systems in support of NATO expeditionary operations. Military communications conference – MILCOM, Baltimore, 2011, 2042–2049.

[3] Bastos L., Wietgrefe H.: Tactical troposcatter applications in challenging climate zones. Military communications conference – MILCOM, Orlando, 2012, 1–6.

[4] Chengcheng X. et al.: Behavioral model measurement system based on signal generator and spectrum analyzer. International Conference on Microwave and Millimeter Wave Technology – ICMMT, 2021 [http://doi.org/10.1109/ICMMT52847.2021.9618281].

[5] Chongjia H. et al.: Photonics-Based Single Sideband Mixer With Ultra-High Carrier and Sideband Suppression. IEEE Photonics Journal 13(4), 2021.

[6] COMET – Compact Over-the-Horizon Transportable Terminal. https://www.comtech.com/comet-compact-over-the-horizon-transportable-terminal (available 24.09.2023).

[7] COMET – CS67PLUS. https://www.comtech.com/cs67plus-troposcatter-modem (available 24.09.2023).

[8] Hao C. et al.: All-Optical In-Phase/Quadrature Microwave Mixer for Antenna Remoting Applications. IEEE Photonics Journal 13(5), 2021, 1–7 [http://doi.org/10.1109/JPHOT.2021.3110589].

[9] Hao C. et al.: Microwave Photonic I/Q Mixer With Phase Shifting Ability. IEEE Photonics Journal 13(4), 2021 [http://doi.org/10.1109/JPHOT.2021.3103786].

[10] Pochernyaev V., Syvkova N.: Broadband switch on partially filled by dielectric rectangular waveguide. The scientific heritage 1(60), 2021, 49–52 [http://doi.org/10.24412/9215-0365-2021-60-1-49-52].

[11] Pochernyaev V. M., Povkhlib V. S.: Mobile combined microwave telecommunication system. 5 International Scientific-Practical Conference Problems of Infocommunications. Scientific – Practical Conference, Kharkiv, 2018.

[12] Pochernyaev V. N., Povkhlib V. S.: Status and directions of development of mobile digital tropospheric communication systems. Systems of arms and military equipment 2(54), 2018, 51–60.

[13] Pochernyaev V. N., Povkhlb V. S.: Status and directions of development of mobile digital radiorelay systems. Systems of arms and military equipment 1(53), 2018, 183–188.

[14] Pochernyayev V. M., Tsibizov K. N.: Complex waveguide theory. Scientific world, Kiev 2003.

[15] Shilpa M. et al.: Low Noise Image-Rejection Gilbert Mixer for Software Defined Radios. IEEE International Conference on Electronics, Computing and Communication Technologies – CONECCT, Bangalore 2020, 1–6 [http://doi.org/10.1109/CONECCT50063.2020.9198584].

[16] Shilpa M. et al.: Recent advancement in the design of mixers for software-defined radios. International Journal of RF and Microwave Computer-Aided Engineering 32(2), 2022, 1–20.

[17] TALON Military Ka Band Terminal. https://www.comtech.com/talon-gx-terminal (available 24.09.2023).

[18] Troposcatter and SATCOM Solutions. https://www. comtech.com/troposcatter-and-satcom-solutions (available 24.09.2023).

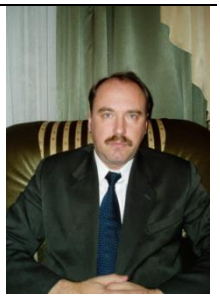
[19] Troposcatter Antenna CSA2400. https://www.comtech.com/troposcatter-antenna, (available 24.09.2023).

[20] Troposcatter Solutions https://www.raytheonintelligenceandspace.com/what-we-do/communications-and-navigation/battlefield-comms/troposcatter (available 24.09.2023).

Prof. Vitaly Pochernyaev
e-mail: vpochernyaev@gmail.com

Doctor of Technical Sciences, professor, member of the Institute of Electrical and Electronics Engineers, professor at the National Academy of Security Service of Ukraine. Author of more than 200 publications, including 3 monographs, 2 textbooks, 12 patents for inventions.
Research interests: applied electrodynamics, radiophysics, telecommunications and radio-engineering, microwave theory and technology.

http://orcid.org/0000-0001-7130-8668



Ph.D. Nataliia Syvkova
e-mail: natsivonat@gmail.com

Doctor of Philosophy, docent National Academy of Security Service of Ukraine. Author of more than 40 publications, including 1 textbook and 4 patents for inventions.
Research interests: telecommunications and radio-engineering, microwave theory and technology.

http://orcid.org/0000-0002-4934-4109

M.Sc. Mariia Mahomedova
e-mail: kkz.praktika@ukr.net

Ph.D. student. Lecturer at the Kyiv Professional College of Communications. Author of more than 20 publications.
Research interests: telecommunications and radio-engineering, microwave theory and technology.

http://orcid.org/0000-0003-1936-5555

http://doi.org/10.35784/iapgos.5394 *received: 09.08.2023 | revised: 01.10.2023 | accepted: 21.11.2023 | available online: 20.12.2023*

INFORMATION SYSTEM FOR DIAGNOSTIC COMPETITIVENESS OF THE HOSPITALITY INDUSTRY OF THE REGIONS OF UKRAINE

Liudmyla Matviichuk[1], Olena Liutak[2], Yuliia Dashchuk[1], Mykhailo Lepkiy[1], Svitlana Sidoruk[1]
[1]Lutsk National Technical University, Faculty of Customs Affairs, Materials and Technologies, Lutsk, Ukraine, [2]Lutsk National Technical University, Faculty of Business and Law, Lutsk, Ukraine

Abstract*. The purpose of this article is to study the main problems and prospects of ensuring the competitiveness of the hospitality industry of the regions of Ukraine in modern conditions, taking into account international experience in the context of deepening integration ties. The work carried out a diagnosis of the level of competitiveness of the hospitality industry of the regions of Ukraine, based on the developed information system of indicators for assessing the conditions of the competitiveness of the hospitality industry of the region and the formed matrix of the competitiveness of the hospitality industry of the region. A comparison was made of the conditions for ensuring competitiveness, the level of competitive advantages and the level of competitiveness of the hospitality industry of the regions of the state in the pre-war period, as well as the diagnosis of the competitiveness of the hospitality industry of the regions of the state was carried out. The positions of the regions in the pre-war and war periods in terms of the level of competitiveness and availability of tourism potential were determined. A matrix for the selection of target indicators of the integration strategy of regions of Ukraine that have preserved their tourist potential is proposed. The results of the study revealed the main problems of ensuring the competitiveness of the hospitality industry in the regions of Ukraine, and highlighted the potential prospects of the studied processes taking into account the conditions of European integration.*

Keywords: information system, competitiveness, hospitality industry

SYSTEM INFORMATYCZNY DLA DIAGNOSTYKI KONKURENCYJNOŚCI BRANŻY GOSPODARCZEJ REGIONÓW UKRAINY

Streszczenie*. Celem artykułu jest zbadanie głównych problemów i perspektyw zapewnienia konkurencyjności branży hotelarskiej regionów Ukrainy we współczesnych warunkach, z uwzględnieniem doświadczeń międzynarodowych w kontekście pogłębiania więzi integracyjnych. W pracy przeprowadzono diagnozę poziomu konkurencyjności branży hotelarskiej regionów Ukrainy na podstawie opracowanego systemu informacyjnego wskaźników oceny warunków konkurencyjności branży hotelarskiej regionu oraz utworzonej macierzy konkurencyjności branży hotelarskiej regionu. Dokonano porównania warunków zapewnienia konkurencyjności, poziomu przewag konkurencyjnych oraz poziomu konkurencyjności branży hotelarskiej regionów państwa w okresie przedwojennym oraz diagnozy konkurencyjności branży hotelarskiej regionów stanu została przeprowadzona. Określono pozycje regionów w okresie przedwojennym i wojennym pod względem poziomu konkurencyjności i dostępności potencjału turystycznego. Zaproponowano macierz doboru wskaźników docelowych strategii integracji regionów Ukrainy, które zachowały swój potencjał turystyczny. Wyniki badania ujawniły główne problemy zapewnienia konkurencyjności branży hotelarskiej w regionach Ukrainy, a także zwróciły uwagę na potencjalne perspektywy badanych procesów z uwzględnieniem uwarunkowań integracji europejskiej.*

Słowa kluczowe: system informacyjny, konkurencyjność, branża hotelarska

Introduction and literature review

The hospitality industry in the regions of Ukraine is represented by a significant tourist potential, wide specialization, the influence of a number of specific factors, and therefore special attention should be paid to the problems of determining the possibilities of competitive development of the studied industry in the regions based on the formation of an information system for the diagnosis of the studied processes. Each territory has its own special socio-economic and resource potential, competitive advantages, a number of factors that influence the development of the hospitality industry in the regions, accelerating or slowing it down. Today in Ukraine, destructive factors of influence prevail (destruction of infrastructure and a number of historical and cultural monuments in many regions of the country due to Russia's military aggression against Ukraine, the spread of pandemics and epidemics, natural and man-made cataclysms, etc.). However, it is the hospitality industry that can become the driving force of the rapid recovery of the regions, provided the appropriate level of ensuring its competitiveness requires the formation of an information system for the study of these processes.

The study of theoretical, methodological aspects and practical principles of ensuring the competitiveness of the hospitality industry of the regions is the subject of research of a number of scientific schools. Both Ukrainian and foreign practitioners and scientists made a significant contribution to the solution of these issues. In particular, certain aspects of ensuring the competitiveness of products and services of the hospitality industry on the basis of sustainable development are the object of research by Holod A., Honcharenko M., Nikyga O., Yevdoshchenko O. [5], Gorina G. [3], Tkachenko T. [14]. In terms of various types of associations of business entities with the aim of ensuring the competitiveness of the hospitality industry, this issue is investigated in the works of Hoblyk V., Papp V. [4], and Mikula N. [7].

The global level of competitiveness of products and services of the hospitality industry is studied by such scientists as Baula O., Liulak O., Chepurda L., Chepurda A., Zelinska O. [1], Smirnov I. [12], Yukhnovska J. [17]. At the level of national competitiveness of products and services of the hospitality industry, research was carried out by such scientists as Shtuler I. [13], Bosovska M. [2]., Nikolaychuk O. [9], Kaplina T., Kaplina A [6], Shykina O. [11]. The regional level of competitiveness is studied in the works of scientists Nezdoyminov S., Baldzhy M., Kniazhkovska H. [8], Vakhovich I., Smal B. [16].

Accordingly, the presence of an analytical base actualizes the issue of modeling diagnostics of the competitiveness of the hospitality industry in the regions of Ukraine. The basis of the work is the study of the dynamics of indicators reflecting the level of competitiveness of the hospitality industry of the regions of Ukraine in the pre-war period, the use of information modeling for the diagnosis of the researched processes, the substantiation of their impact on the competitiveness of the hospitality industry of the studied territories. These conclusions will make it possible to develop recommendations based on the European practice of forming strategic programs for strengthening the competitiveness of the hospitality industry.

1. Materials and methods

The following research methods were used:

- methods of economic analysis - to analyze and evaluate the conditions for ensuring the competitiveness of the hospitality industry in the region, the competitive advantages of the hospitality industry in the social, economic and environmental spheres;

- the method of standardization of indicators – for the calculation of partial standardized indicators that characterize the conditions for ensuring the competitiveness of the hospitality industry of the region, the competitive advantages of the hospitality industry of the region in the economic, social and environmental spheres;
- the method of integrating indicators – for calculating the integral index of competitiveness of the hospitality industry of the region;
- rating method – to determine the rating of regions according to the level of competitiveness of the hospitality industry of the region;
- method of grouping – for grouping regions according to the level of conditions for ensuring the competitiveness of the hospitality industry of the region; the level of competitive advantages of the hospitality industry in the social, economic and environmental spheres; the level of competitiveness of the region's hospitality industry;
- graphic methods – for visual representation of the results of an analytical study of the competitiveness of the hospitality industry in the region.

The primary data for analysis were statistical data from UNWTO [10, 15].

After collecting and processing statistical data, a diagnosis of the competitiveness of the hospitality industry in the regions of Ukraine was conducted. The collected and processed statistical material of the analysis made it possible to describe the state, structure and dynamics of the relevant indicators of the hospitality industry in the regions of Ukraine. The results of the statistical analysis made it possible to identify problematic points, positive aspects and shortcomings for the formation of a matrix for the selection of target indicators of the integration strategy of regions of Ukraine that have preserved their tourist potential.

2. Results and discussion

Increasing the competitiveness of the hospitality industry in the region requires complex and systematic analytical studies to identify the level, trends and competitive advantages of each territory of the state in relation to its social resources, the existing institutional environment for the development of the hospitality industry, favorable regional conditions for the effective use of tourism potential, etc. [5]. On the other hand, it is important to study the available resource support for the development of the hospitality industry in the regions of the state from the standpoint of the presence of social, economic, tourist and informational resources as prerequisites for ensuring their competitiveness [3]. In connection with this, issues regarding the choice of a balanced system of indicators and an algorithm of actions for the analysis and assessment of the competitiveness of the hospitality industry of the region in the form of an information system are being updated.

The analysis and assessment of the competitive advantages of the hospitality industry of the regions was carried out in the context of three spheres: economic, social and environmental.

The selection of a system of indicators to assess the conditions for ensuring the competitiveness of the hospitality industry of the region and the competitive advantages of the hospitality industry of the region in the economic, social and environmental spheres (table 1).

In the course of the study, the matrix of the indicator system was obtained, which reflects the conditions for ensuring the competitiveness of the hospitality industry and the competitive advantages of the hospitality industry of the region in the economic, social and environmental spheres.

Table 1. A system of indicators for assessing the conditions of competitiveness of the hospitality industry of the region (compiled by the author based on: [1, 6, 11, 12, 14, 16])

A group of indicators characterizing	Indexes
Resource conditions for ensuring the competitiveness of the hospitality industry of the region	Available population, total fertility rate, labor force, employed local population, proportion of households with internet access, area of protected land and national natural parks
Institutional conditions for ensuring the competitiveness of the hospitality industry in the region	The number of children's health and recreation facilities, the number of places in children's recreation and health facilities, the numer of collective accommodation facilities, the number of tourism entities, the number of tour operators, the number of travel agents, the number of museums
Geopolitical conditions for ensuring the competitiveness of hospitality in the region	The number of employed population in the hospitality industry, the level of GVA of the hospitality industry in the structure of the economy

Such a matrix looks as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{in} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{w1} & x_{w2} & \cdots & x_{wk} & \cdots & x_{wn} \end{bmatrix}$$

here w is the number of regions; n – the number of indicators that characterize the conditions for ensuring the competitiveness of the hospitality industry and the competitive advantages of the hospitality industry of the region in the economic, social and environmental spheres; X_{ik} is the value of the indicator k for the i-th region.

The given system of indicators made it possible to standardize them and calculate an integral index of conditions for ensuring the competitiveness of the region's hospitality industry and competitive advantages of the region's hospitality industry by sphere.

The standardization of indicators was carried out with the aim of bringing the indicators, which are different in terms of content and measurement units, to a single coefficient, the calculation of which will allow to further determine the integral index of the conditions for ensuring the competitiveness of the hospitality industry and the competitive advantages of the hospitality industry of the region in the economic, social and environmental spheres.

Standardization involves the calculation of stimulator indicators (those that have a positive effect on the researched processes) and destimulant indicators (those that negatively affect the researched processes) of the competitiveness of the hospitality industry:

$$X_{ij}^{I} = \frac{X_{ij}}{X_{m+1j}} \quad (1)$$

or

$$X_{ij}^{I} = \frac{X_{m+1j}}{X_{ij}} \quad (2)$$

here $i = 1, 2, ..., m$, $j = 1, 2, ..., n$, j-th is the indicator of the best (reference) region (reference is a region with high values of indicators of the studied processes). At the same time, formula (1) is used when the reference indicator is the maximum value, and formula (2) is the minimum value.

At the next stage, the integral index of competitiveness of the hospitality industry of the region was calculated. At the same time, the integral index of the conditions for ensuring the competitiveness of the hospitality industry of the region as calculated as the arithmetic average of three components that characterize it:

$$IIc = \frac{Rc+Ic+Gc}{3} \quad (3)$$

IIc is an integral index of the conditions for ensuring the competitiveness of the hospitality industry of the region; *Rc* – resource conditions for ensuring the competitiveness of the hospitality industry of the region; *Ic* – institutional conditions for ensuring the competitiveness of the hospitality industry of the region; *Gc* – geopolitical conditions for ensuring the competitiveness of the hospitality industry in the region.

The integral index of the competitive advantages of the hospitality industry of the region was calculated as the arithmetic average of three spheres:

$$\text{Iic} = \frac{\text{Esp}+Ssp+ECsp}{3} \quad (4)$$

Iic is an integral index of the competitive advantages of the hospitality industry of the region; *Esp* – the economic sphere of the hospitality industry of the region; *Ssp* – is the social sphere of the region's hospitality industry; *ECsp* – is the ecological sphere of the region's hospitality industry.

Thus, the integral index of competitiveness of the hospitality industry of the region was calculated according to the following formula:

$$Ict = \frac{Ic+Ica}{2} \quad (5)$$

Ict – is an integral index of competitiveness of the hospitality industry of the region; *Ic* – is an integral index of the conditions for ensuring the competitiveness of the hospitality industry of the region; *Ica* – is an integral index of the competitive advantages of the hospitality industry of the region.

The calculated integral index of the competitiveness of the hospitality industry of the region may vary within [0; 1]. The closer the value of the index is to 1, the more competitive the hospitality industry will be in the region, and vice versa – the closer it is to 0, the region will be characterized by a low level of competitiveness of the hospitality industry.

To establish the limits of the intervals regarding the competitiveness of the hospitality industry of the region, we offer the following levels:

– for the conditions of ensuring the competitiveness of the hospitality industry of the region:

0.00–0.49 – low level of conditions for ensuring the competitiveness of the hospitality industry of the region;

0.50–1.00 – a high level of conditions for ensuring the competitiveness of the hospitality industry in the region;

– for the competitive advantages of the region's hospitality industry in the social, economic and environmental spheres:

0.00–0.49 – low level of competitive advantages of the hospitality industry of the region in the social, economic and environmental spheres;

0.50–1.00 – high level of competitive advantages of the hospitality industry of the region in the social, economic and environmental spheres;

– for the competitiveness of the region's hospitality industry:

0.00–0.33 – low level of competitiveness of the hospitality industry of the region;

0.34–0.67 – the average level of competitiveness of the hospitality industry of the region;

0.68–1.00 – high level of competitiveness of the hospitality industry of the region.

At the last stage, a matrix of competitiveness of the hospitality industry of the region was formed (Fig. 1).

As we can see from the table, the economic, according to the matrix, we obtained eight types of ratios according to the level of competitiveness of the hospitality industry of the region:

– 1 type of regions: l/l/l – low level of conditions for ensuring the competitiveness of the region; low level of competitive advantages of the hospitality industry of the region; low level of competitiveness of the region's hospitality industry;
– 2 type of regions: l/l/m – low level of conditions for ensuring the competitiveness of the region; low level of competitive advantages of the hospitality industry of the region; average level of competitiveness of the region's hospitality industry;
– 3 type of regions: h/l/l – low level of conditions for ensuring the competitiveness of the region; low level of competitive advantages of the hospitality industry of the region; high level of competitiveness of the region's hospitality industry;
– 4 type of regions: l/h/m – low level of conditions for ensuring the competitiveness of the region; high level of competitive advantages of the region's hospitality industry; average level of competitiveness of the region's hospitality industry;
– 5 type of regions: l/h/h – low level of conditions for ensuring the competitiveness of the region; high level of competitive advantages of the region's hospitality industry; high level of competitiveness of the region's hospitality industry;
– 6 type of regions: h/l/l – high level of conditions for ensuring the competitiveness of the region; low level of competitive advantages of the hospitality industry of the region; low level of competitiveness of the region's hospitality industry;
– 7 type of regions: h/l/m – high level of conditions for ensuring the competitiveness of the region; low level of competitive advantages of the hospitality industry of the region; average level of competitiveness of the region's hospitality industry;
– 8 type of regions: h/h/m – high level of conditions for ensuring the competitiveness of the region; high level of competitive advantages of the region's hospitality industry; average level of competitiveness of the region's hospitality industry;
– 9 type of regions: h/h/h – high level of conditions for ensuring the competitiveness of the region; high level of competitive advantages of the region's hospitality industry; high level of competitiveness of the region's hospitality industry.

Each type of region requires a differentiated approach to choosing a strategy to ensure the competitiveness of the region's hospitality industry.

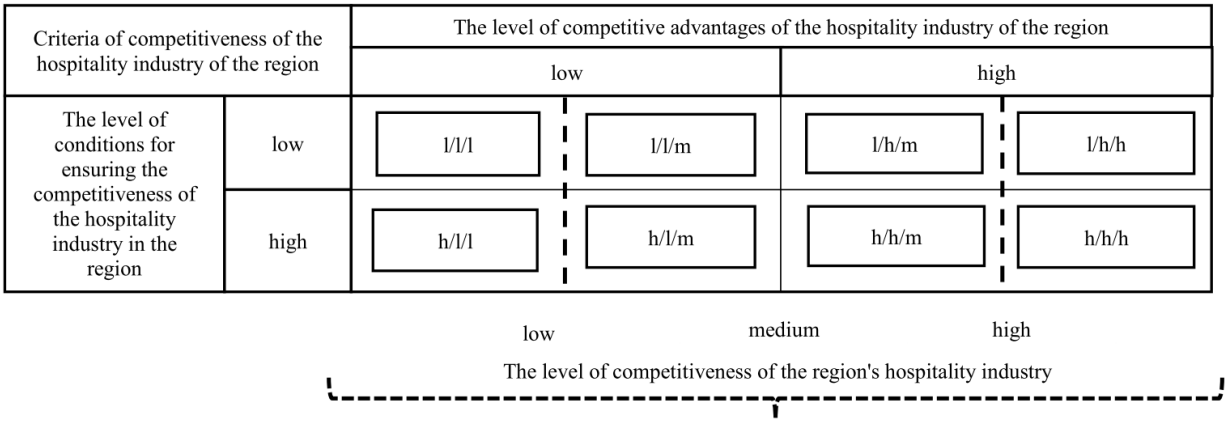


Fig. 1. Matrix of competitiveness of the hospitality industry of the region

3. Experiment

The next step in our research was the calculation of the level of competitiveness of the hospitality industry of the regions of Ukraine, according to the formed information system. The initial data for building this system is presented in table 4.

The comparison of the level of conditions for ensuring competitiveness, the level of competitive advantages and the level of competitiveness of the hospitality industry of the regions of the state in 2021 is shown in table 2.

We see that there are different ratios in these levels. The ideal ratio "high level of conditions for ensuring competitiveness – high level of competitive advantages of the hospitality industry – high level of competitiveness of the hospitality industry" is typical for the city of Kyiv. In this region, a high level of competitiveness of the hospitality industry is observed due to the formed and effectively used conditions for ensuring such competitiveness, as well as due to the presence of competitive advantages in the field of the hospitality industry compared to other territories of the state. Dnipropetrovsk, Zakarpattia, Lviv, Odesa, Ternopil, Kharkiv and Chernivtsi regions are characterized by the ratio "low/high level of conditions for ensuring competitiveness – low/high level of competitive advantages of the hospitality industry – low/high level of competitiveness of the hospitality industry". Such discrepancies indicate ineffective use by the territory of its potential opportunities (conditions) and unformed competitive advantages that ensure the competitiveness of the hospitality industry in the regions of the state. The rest of the regions have the ratio "low level of conditions for ensuring competitiveness – low level of competitive advantages of the hospitality industry – low level of competitiveness of the hospitality industry." These regions (Volhynia, Vinnytsia, Donetsk, Zhytomyr, Zaporizhia, Kyiv, Kirovohrad, Ivano-Frankivsk, Luhansk, Poltava, Rivne, Mykolaiv, Ternopil, Sumy, Khmelnytsky, Kherson, Chernihiv, Cherkasy) are outsiders in ensuring the competitiveness of the hospitality industry, which requires the development and implementation of a stabilization strategy through the implementation of a system of adaptive measures to eliminate problems in the hospitality industry.

As a result of the conducted analytical studies, a diagnosis of the competitiveness of the hospitality industry in the regions of the state was made (table 3).

Table 2. Comparison of the level of conditions for ensuring competitiveness, the level of competitive advantages and the level of competitiveness of the hospitality industry of the regions of the state in 2021 in the pre-war period (calculated by the authors)

Regions	Criteria index	Level	Competitive Advantage Index	Level	Competitiveness index	Level
Vinnytsia	0.250	low	0.185	low	0.218	low
Volynska	0.300		0.325		0.313	
Dnipropetrovsk	0.582	high	0.210		0.396	
Donetsk	0.329	low	0.251		0.290	
Zhytomyr	0.272		0.341		0.307	
Zakarpatia	0.409		0.584	high	0.497	high
Zaporizka	0.441		0.239	low	0.340	low
Ivano-Frankivsk	0.338		0.191		0.265	
Kyivska	0.442		0.198		0.320	
Kirovograd	0.225		0.225		0.225	
Lugansk	0.145		0.280		0.213	
Lviv	0.645	high	0.243		0.444	high
Mykolaiv	0.321	low	0.309		0.315	low
Odessa	0.639	high	0.337		0.488	high
Poltava	0.324	low	0.174		0.249	low
Rivne	0.309		0.212		0.261	
Sumska	0.214		0.162		0.188	
Ternopilska	0.268		0.425	high	0.347	
Kharkiv	0.505	high	0.162	low	0.334	
Kherson	0.319	low	0.266		0.293	
Khmelnytskyi	0.312		0.154		0.233	
Cherkasy	0.346		0.171		0.259	
Chernivtsi	0.269		0.748	high	0.509	high
Chernihivt	0.239		0.185	low	0.212	low
m. Kyiv	0.864	high	0.488	high	0.676	high

Table 3. Diagnostics of competitiveness of the hospitality industry of the regions of the state in the pre-war period (compiled according to [9, 13, 12])

Type of regions	Leading regions	Middle regions	Outsider regions
Regions	city Kyiv	Zakarpattia, Lviv, Odesa, Chernivtsi	Volyn, Vinnytsia, Dnipropetrovsk, Donetsk, Zhytomyr, Zaporizhia, Kyiv, Kirovohrad, Ivano-Frankivsk, Luhansk, Mykolayiv, Rivne, Poltava, Ternopil, Sumy, Kharkiv, Kherson, Khmelnytsky, Cherkasy, Chernihiv Chernivtsi
Growth rate	More positive	Positive/negative	Mostly negative
Strengths (advantages)	A large area of lands of reserves and national nature parks, developed institutional infrastructure of the hospitality industry, high tourist potential, availability of tourist flows, investment attractiveness for the development of hospitality industry facilities, developed information network of familiarization with services in the hospitality industry, favorable environmental situation	Provision of natural and recreational resources, awareness of the population with tourist products, stable income from tourism activities, implementation of a system of measures for waste disposal	The presence of positive growth rates for certain indicators characterizing the competitiveness of the hospitality industry (the share of households that have access to the Internet, the volume of output, the number of museums, average monthly wages, tourist tax)
Weaknesses (problems)	The presence of negative growth rates on certain indicators characterizing the competitiveness of the hospitality industry (population, labor force, number of tourism entities, number of KZR, number of people who were in KZR, number of actually spent bed days in KZR)	Lack of qualified force in the hospitality industry, insufficient level of investment attractiveness of the territory, low pay for hospitality workers, insufficient number of people in the KZR, decrease in the number of labor force, decrease in the number of tourism entities	Low output in the hospitality industry, low labor cost, low employment, morally and economically obsolete technologies for providing services in the hospitality industry, unsatisfactory ecological state of territories, low level of access to the Internet, inefficient use of the territory's potential for the development of the hospitality industry, low investment in hospitality industry
Possibilities of the external environment	Improving the business climate, stimulating the development of the hospitality industry, increasing the popularity of domestic tourism, increasing the investment attractiveness of territories, forming environmental awareness of the population, activating grant and project work in the field of the hospitality industry		
Threats of the external environment	Unfavorable economic, political and environmental situation in the country and regions in particular, migration of skilled labor in the field of tourism, decrease in the purchasing power of the population, decrease in the financial capacity of territories, inability to implement modern technologies in the field of hospitality industry		

Therefore, each group of regions is characterized by different rates of growth, the presence of opportunities and threats, strengths and weaknesses that shape the external environment. For the leading regions (the city of Kyiv), the strengths in ensuring the competitiveness of the hospitality industry are the large area of land reserves and national natural parks, the developed institutional infrastructure of the hospitality industry, high tourist potential, the presence of tourist flows, investment attractiveness for the development of hospitality industry facilities, a developed information network for familiarization with services in the hospitality industry, a favorable environmental situation. At the same time, the regions have weaknesses, which in the future may weaken their position in ensuring the competitiveness of the hospitality industry: a decrease in the population, a decrease in the number of the workforce, a decrease in the number of subjects of tourist activity, a decrease in the number of collective means of accommodation, a decrease in the number of persons who stayed in them, a decrease in the number of bed-days actually spent in collective means of accommodation). As we can see, the regions against the backdrop of growth rates for most indicators also have negative growth trends for some indicators, which requires the use of preventive measures to minimize and eliminate them.

Strengths for regions that have different ratios in the levels of competitiveness of the hospitality industry are the supply of natural and recreational resources, the population's awareness of tourist products, stable income from tourism activities, and the implementation of a system of waste disposal measures. On the contrary, their weaknesses that do not allow to achieve a high level of conditions for ensuring the competitiveness of the hospitality industry, a high level of competitive advantages of the hospitality industry, a high level of competitiveness of the hospitality industry are the lack of qualified personnel in the field of the hospitality industry, insufficient level of investment attractiveness of the territory, low wages of hospitality industry workers, an insufficient number of people staying in collective means of accommodation, a decrease in the number of the workforce, a decrease in the number of subjects of tourist activity. The mentioned problems should become the object of research for regional authorities and subjects of tourism activity.

Outsider regions have significant problems that form their weak points in ensuring the competitiveness of the hospitality industry: low output in the hospitality industry, low labor costs, low employment of the population, morally and economically outdated technologies for providing services in the hospitality industry, unsatisfactory ecological condition of the territories, low level of access to the Internet, inefficient use of the potential of the territory for the development of the hospitality industry, low volume of investment in the facilities of the hospitality industry. However, the presence of advantages in growth rates (the share of households with access to the Internet, the volume of production, the number of museums, average monthly wages, and tourist tax) will allow these regions to stabilize the situation and create favorable conditions and competitive advantages in ensuring the competitiveness of the hospitality industry.

Authorities at the state and regional levels, the institutional environment create opportunities (improving the business climate, stimulating the development of the hospitality industry, increasing the popularity of domestic tourism, increasing the investment attractiveness of territories, forming environmental awareness of the population, activating grant and project work in the field of hospitality industry) and threats (unfavorable economic, political and environmental situation in the country and regions in particular, migration of skilled labor in the field of tourism, decrease in the purchasing power of the population, decrease in the financial capacity of territories, inability to implement modern technologies in the field of hospitality industry) to each region in ensuring the competitiveness of the hospitality industry. Therefore, regions should make the most of their strengths and opportunities created by the external environment to strengthen their competitive positions in terms of the development of the hospitality industry.

It is worth noting that the presence of military actions in Ukraine since 2022 has changed the conditions for ensuring the competitiveness of certain regions of the state and their competitive advantages in the economic, social and environmental spheres. The tourism potential of many regions has been destroyed, the subjects of the tourism industry have stopped their activities due to the impossibility of carrying them out due to danger and threats, and a large part of the population has emigrated abroad or to safer regions. All this has a negative impact on the competitiveness of the hospitality industry in certain regions of the state and the country as a whole. Therefore, it is worth highlighting the regions that are currently in the zone of active hostilities and have lost their tourist potential, and that need special state support in the post-war period (table 4).

Table 4. Positioning of regions in the pre-war and post-war periods (calculated by the authors)

Positioning of regions in the pre-war period	
Low level of competitiveness of the hospitality industry	High level of competitiveness of the hospitality industry
Volyn, Vinnytsia, Dnipropetrovsk, Donetsk, Zhytomyr, Zaporizhia, Kyiv, Kirovohrad, Ivano-Frankivsk, Luhansk, Mykolayiv, Rivne, Poltava, Ternopil, Sumy, Kharkiv, Kherson, Khmelnytsky, Cherkasy, Chernihiv.	Zakarpattia, Lviv, Odesa, Chernivtsi, Kyiv.
Positioning of regions during the war period	
Regions that have lost their tourist potential	Regions that have preserved tourist potential
Dnipropetrovsk, Donetsk, Zaporizhzhya, Kyiv, Luhansk, Mykolaiv, Sumy, Ternopil, Kharkiv, Kherson, Chernihiv, Kyiv.	Vinnytsia, Volyn, Zakarpattia, Zhytomyr, Lviv, Odesa, Ivano-Frankivsk, Kirovohrad, Poltava, Rivne, Khmelnytskyi, Cherkasy, Chernivtsi.

For regions that have lost their tourist potential (Donetsk, Dnipropetrovsk, Zaporizhzhia, Kyiv, Luhansk, Mykolaiv, Sumy, Ternopil, Kharkiv, Kherson, Chernihiv, Kyiv) a balanced state and regional policy aimed at restoring conditions and competitive advantages should be developed in ensuring the competitiveness of the hospitality industry. Thus, the typology of regions of the state according to the level of competitiveness of the hospitality industry in the pre-war and post-war periods requires the use of a differentiated approach to the development of a strategy for its improvement.

When forming the integration strategy of the hospitality industry of the region, one should take into account the peculiarities of the regions of Ukraine, as potential participants of international economic groups (cross-border clusters, strategic alliances of the hospitality industry, etc.) [4, 7]. A clearly defined information system of integration elements (policies, programs, techniques, methods, methods, tools, etc.) will allow speeding up the achievement of success. However, ensuring the competitiveness of the hospitality industry of the regions is not a stable process, but a dynamic one and requires constant control [10]. Considering this, it is advisable to form target indicators of the integration strategy for each defined type of regions of Ukraine for a general system of indicators, which will facilitate management processes in terms of control in table 5.

The expected result of the implementation of integration strategies in the regions of Ukraine is the strengthening of the competitiveness of the hospitality industry [2, 17]. The effectiveness of the implementation of these strategies can be determined with the help of target indicators, that is, clearly defined indicators, with the help of which conclusions can be drawn about the success of the implementation of the selected integration strategy in the region. The target indicators of the integration strategy of Ukraine's regions defined by us are measurable, which greatly simplifies the control function.

Table 5. Matrix for the selection of target indicators of the integration strategy of regions of Ukraine that have preserved their tourist potential (calculated by the authors)

Types of regions	Groups of regions in the selection matrix of target indicators	Integration strategies	Target indicators	
			Regional	General
Regional leaders	city Kyiv	Integration transformation strategy	– Creation of cross-border hospitality cooperation. – Ensuring the functioning of the hospitality industry in accordance with international standards – Development of competencies of representatives of the hospitality industry	– Sustainable growth of quantity, quality and competitive hospitality products and services at the national, regional and local levels. – Increasing Ukraine's position in the Global Competitiveness Index – Increasing the share of industries in the added value of the country – Creation of a competitive hospitality industry market. – Development of hospitality industry infrastructure. – Improving the quality of hospitality products and services. – Accelerating the integration of the hospitality industry into the global services market. – Increasing the number of associations of hospitality industry subjects. – Increasing the number of internal and external investments. – Increasing the number of geographical indications in the regions. – Improvement of socio-economic indicators of the hospitality industry in the region. – Increasing the share of industry in GDP to the world average level. – Stimulating the development of hospitality innovations.
Middle peasant regions	Odesa, Zakarpattia, Lviv	Integration development strategies	– Formation of high-quality and comprehensive hospitality products and services. – Increasing the number of associations of hospitality entities. – Stimulation of investments in projects related to improving the condition of tourist destinations. – Increasing the share of small and medium-sized hospitality businesses	
Outsider regions	Vinnytsia, Volyn, Zhytomyr, Ivano-Frankivsk, Kirovohrad, Poltava, Rivne, Khmelnytsky, Cherkasy, Chernivtsi	Integration interaction strategy	– Formation of new and improvement of existing hospitality industry clusters and other associations. – Increasing the volume of tourist flows – Increasing the number of people employed in the hospitality industry to the total number of employed. – Improvement of existing and formation of new regional quality standards.	

The strengthening of the integration processes of business entities of the hospitality industry is based on the existing market structure in the region, specific features of the industry of the regions, the intensity of consumption of the offered hospitality products and services, the level of application of innovations in the researched field, etc. [8, 11]. Today, the integration processes of the hospitality industry depend on the specifics of the relationship with other business entities related to hospitality. In this case, the research processes are formed based on international competition and market concentration of hospitality industry resources. In addition, before starting regional strategic planning for the development of the researched industry, it is necessary to take into account the specifics of the development of the hospitality industry.

A mandatory condition for the effective development of the modern hospitality industry is the coherence of strategic decisions at all hierarchical levels, for this purpose a hierarchy of strategic relations of the hospitality industry was built, where four main levels are distinguished: global or international, national, regional and local or local. It is justified that normative documents should include clear and specific information, normative legal documents of state regulation of hospitality should take into account the interests of all stakeholders (business entities, enterprises and related industries) operating in the market. Territorial development programs, regional strategies and other regulatory documents for the development of the industry under study must comply with international standards, norms and rules, as well as be based on the interests of regional (small and medium-sized) businesses, take into account the goals of national strategic decisions. In addition, strategic decisions at the level of local business entities must take into account the interests of the state and the need to protect and reproduce the historical, cultural and natural resources of the region, the state and regional concept of market positioning.

4. Conclusions

An information system and a methodical approach to the analysis and assessment of the competitiveness of the hospitality industry of the region have been developed, which allows to assess the resource, institutional, geopolitical conditions for ensuring the competitiveness of the hospitality industry and its competitive advantages in the economic, social and environmental spheres, and on this basis – to calculate the integrated index of the industry's competitiveness hospitality

The analysis and assessment of the competitiveness of the hospitality industry of the regions of the state in terms of the conditions for ensuring the competitiveness of the hospitality industry and its competitive advantages in the social, economic and environmental spheres was carried out. The diagnosis of the competitiveness of the hospitality industry in the regions of the state was carried out. The integral index of the conditions for ensuring the competitiveness of the hospitality industry of the regions of the state, the integral index of the competitive advantages of the hospitality industry of the regions of the state, the integral index of the competitiveness of the hospitality industry of the regions of the state was calculated.

Each group of regions (leading regions, middle-class regions, outsider regions) is characterized by different growth rates, the presence of strengths and weaknesses, opportunities and threats that shape the external environment. For the leading regions (the city of Kyiv), the strengths in ensuring the competitiveness of the hospitality industry are the large area of land reserves and national natural parks, the developed institutional infrastructure of the hospitality industry, high tourist potential, the presence of tourist flows, investment attractiveness for the development of hospitality industry facilities, a developed information network for familiarization with services in the hospitality industry, a favorable environmental situation. At the same time, these regions have weaknesses that in the future may weaken their position in ensuring the competitiveness of the hospitality industry: a decrease in the population, a decrease in the number of the workforce, a decrease in the number of subjects of tourist activity, a decrease in the number of collective means of accommodation, a decrease in the number of people who stayed with them, a decrease in the number of bed-days actually spent in collective means of accommodation).

For regions that have lost their tourist potential (Donetsk, Dnipropetrovsk, Zaporizhzhia, Kyiv, Luhansk, Mykolaiv, Sumy, Ternopil, Kharkiv, Kherson, Chernihiv, Kyiv) it is advisable to develop a balanced state and regional policy aimed at restoring conditions and competitive advantages in ensuring the competitiveness of the hospitality industry. Thus, the typology of regions of the state according to the level of competitiveness of the hospitality industry in the pre-war and post-war periods requires the use of a differentiated approach to the development of a strategy for its improvement.

The process of formation of strategic integration positions of the regions as a set of stages is substantiated, which reflects the logic of the formation of the algorithm of actions through the definition of structural elements that ensure the formation of an effective toolkit for achieving the conceptual goal. The algorithm for choosing the strategic integration positions of the regions should be complex and systemic, and all strategic elements of the integration processes of the region should be coordinated and aimed at achieving the strategic goal – the definition of an effective, optimal strategy that is able to ensure the implementation of the target priorities and strategic guidelines of the proposed concept of ensuring competitiveness hospitality industry of the region.

A matrix for choosing the type of strategies groups of regions has been built, which allows for effective management of the hospitality industry in the region, applying a system of functions, principles, tools and methods for ensuring the competitiveness of the region's hospitality, taking into account the influence of modern factors. The proposed matrix is adaptive, able to modify the structure of the hospitality industry in the region due to the formation of new elements (associations of hospitality entities), the appearance of which is conditioned by the spectrum of defined target tasks.

References

[1] Baula O., Liulak O., Chepurda L., Chepurda A., Zelinska O.: Financial levels and consequences of implementation of world experience of business activity in domestic tourism. Financial and credit activity: problems of theory and practice. University of Banking 4 (35), 2020, 476–485.

[2] Bosovska M.: Integration processes in tourism: monograph. National Trade and Economy Univ., Kyiv 2015.

[3] Gorina G.: Hierarchical clustering of regional hotel services markets in Ukraine by infrastructure components. Investment: Practice and Experience 23, 2019, 5–10.

[4] Hoblyk V., Papp V.: The formation of cross-border transport and logistics clusters as a priority direction for the development of the border regions of Ukraine. Economic Forum 4, 2013 55–63.

[5] Holod A., Goncharenko M.., Nikyga O., Yevdoshchenko O.: Innovative principles of sustainable development of ethno-gastronomic tourism in the region. Manager. Bulletin of the Donetsk State University of Management. Economy series 4 (89), 2020, 40–47.

[6] Kaplina T., Kaplina A.: Hospitality industry of Ukraine: development trends. Collection of scientific works Problems and prospects of entrepreneurship development 27, 2021, 116–127.

[7] Mikula N.: Priorities for the creation of innovation-investment cross-border clusters and industrial zones in the western regions of Ukraine: analytical note to the Administration of the President of Ukraine, 2011. http://www.niss.lviv.ua/file/2011/03/173222 .pdf (available: 28.06.2023)

[8] Nezdoyminov S., Baldzhy M., Kniazhkovska H.: Regional measurement of the hotel sector development of a tourist destination (on the example of Odessa region, Ukraine). Journal of Geology, Geography and Geoecology 30 (2), 2021, 344–354.

[9] Nikolaychuk O.: Development trends of the domestic hospitality industry in the conditions of Covid-19. Bulletin of KhNU named after Karazin. Series "International relations. Economy. Local studies. Tourism" 13, 2021.

[10] Prychepa I., Adler O., Ruda L., Lesko O., Bondarenko Z., Yanan L., Mussayeva D.: An information model for assessing the competitiveness of the tourism industry in the context of European integration policy. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska – IAPGOS 12(4), 2022, 47–52.

[11] Shikina O.: Index of tourism competitiveness as an indicator of tourism potential of Ukraine. Bulletin of the Khmelnytskyi National University 2, 2017, 275–279.

[12] Smirnov I.: Determining the level of competitiveness of the countries of the world in tourism: modern approaches. Foreign trade: economy, finance, law. Series: Economic Sciences 5-6(82-83), 2015, 37–49.

[13] Stuler I.: Peculiarities of the development of the hospitality industry in the conditions of the European integration vector of Ukraine. Actual problems of the economy 1, 2022, 79–86.

[14] Tkachenko T.: Competitiveness in tourism as a leading direction of sustainable development. Scientific and information bulletin of the Academy of Sciences of Higher Education of Ukraine 2(79), 2012, 96–104.

[15] UNWTO, Tourism Highlights, http://www2.unwto.org (available: 28.06.2023)

[16] Vakhovich I., Matviychuk L., Smal B.: The development of the hospitality industry in modern conditions: trends and measures to strengthen competitive advantages. Financial and credit activity: problems of theory and practice 6(41), 2021, 494–502 [http://doi.org/10.18371/fcaptp.v6i41.251512].

[17] Yukhnovska J.: Impact of globalization and European integration on the tourism industry of Ukraine. Scientific Bulletin of the Uzhhorod National University. Series: International economic relations and the world economy 23, 2019, 147–152.

D.Sc. Liudmyla Matviichuk
e-mail: kalishl.y@gmail.com

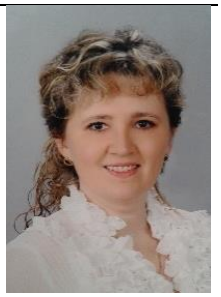
Doctor of Economics, Professor. Scientific direction: rational use of tourist resources, competitiveness of the hospitality industry, quality assurance of hospitality products and services, integration processes of the hospitality industry.

http://orcid.org/0000-0003-1694-6178

D.Sc. Olena Liutak
e-mail: olenalutak@gmail.com

Doctor of Economics, Professor. Scientific direction: global institutionalization of the service sector, international competitiveness, international tourism, regulation of the development of global food and non-food markets; innovation and investment potential.

http://orcid.org/0000-0002-4293-0586



Ph.D. Yuliia Dashchuk
e-mail: yli2204@ukr.net

Scientific direction: regional economics, sustainable tourism, cultural heritage, tourism policy.

http://orcid.org/0000-0001-6394-017X

Ph.D. Mykhailo Lepkiy
e-mail: lepkiym@gmail.com

Scientific direction: geography of tourism, organization of tourism, gastronomic tourism, tourist resources of Ukraine and the world, information technologies of the hospitality industry.

http://orcid.org/0000-0003-2470-6780

Ph.D. Svitlana Sidoruk
e-mail: sidoruk_svitlana@ukr.net

Scientific direction: regional development, tourism business, hotel and restaurant potential of the region, community branding, inclusive tourism.

http://orcid.org/0000-0002-7403-6909

http://doi.org/10.35784/iapgos.5604 *received: 10.10.2023 | revised: 04.11.2023 | accepted: 24.11.2023 | available online: 20.12.2023*

ENVIRONMENTAL AND ECONOMIC ASSESSMENT OF THE LAND USE REGULATION EFFECTIVENESS

Oleksandr Harnaha[1], Natalia Savina[1], Volodymyr Hrytsiuk[2]
[1]National University of Water and Environmental Engineering, Institute of Economics and Management, Rivne, Ukraine, [2]National University of Water and Environmental Engineering, Nadsluchansky Institute, Berezne, Ukraine

Abstract. *The article reveals the aspects of environmental and economic assessment of the land use regulation effectiveness. It is emphasized that in order to use land resources rationally and successfully implement scientifically sound farming methods, it is necessary to have holistic and reliable information about land, its productive properties, natural and economic condition. It is noted that the efficiency of land use should have quantitative parameters characterizing each of the aspects directly involved in the process of land use regulation.*

Keywords: land pollution, soil, economic indicators, environmental economics

ŚRODOWISKOWA I EKONOMICZNA OCENA SKUTECZNOŚCI ROZPORZĄDZENIA O UŻYTKOWANIU GRUNTÓW

Streszczenie. *W artykule przedstawiono aspekty środowiskowej i ekonomicznej oceny skuteczności regulacji dotyczących użytkowania gruntów. Podkreśla się, że aby racjonalnie wykorzystywać zasoby ziemi i skutecznie wdrażać naukowo uzasadnione metody gospodarowania, konieczne jest posiadanie całościowej i rzetelnej informacji o ziemi, jej właściwościach produkcyjnych, warunkach przyrodniczych i ekonomicznych. Należy zauważyć, że efektywność użytkowania gruntów powinna mieć parametry ilościowe charakteryzujące każdy z aspektów bezpośrednio zaangażowanych w proces regulacji użytkowania gruntów.*

Słowa kluczowe: zanieczyszczenie gleby, gleba, wskaźniki ekonomiczne, ekonomika środowiska

Introduction

One of the most important principles of land use is the establishment of a scientifically sound relationship between economic and environmental parameters. This is also an important principle of balance and conformity. Land use is a complex integrated economic system that contains many subsystems. In this regard, it is necessary to take into account the degree of human activity impact on the environment and strictly comply with the requirements of ecological balance in each specific land use in the process of land management. In order to use land resources rationally and successfully implement scientifically sound farming practices, it is vital to have complete and reliable information about the land, its productive properties, natural and economic condition. The point is that the lands of different zones, districts, farms and even individual plots, which have their own special spatial conditions, relief, soil cover, vegetation, water regime, geological structure, as well as other natural and climatic resources, vary greatly in terms of their properties and suitability for agricultural production. Therefore, these land properties have a different impact not only on crop yields, but also on production costs, labor productivity, gross product and net income, as well as on the development and implementation of measures aimed at protection of land and related natural resources and surrounding environment.

Ecological and economic aspects of land use development have been holistically studied by the domestic scientists as H. Hutsuliak [5], D. Dobriak [2], A. Tretiak [10], A. Sokhnych [9] and others.

It has been established that if the land is properly cultivated, it increases its fertility and has a beneficial effect on the environment. At the same time, the study of this issue and numerous data show that while in the process of use, land properties haven't been improved, but deteriorated and even destroyed, if the wrong farming methods are used.

In this regard, there are numerous recommendations to improve technological processes in agricultural technology, preparation and application of fertilizers, and the use of complex mechanization in soil and crop cultivation. At the same time, the issues of technical influence on the structural condition of the soil are not sufficiently studied today. Farms located in different natural and climatic zones with completely different soils in terms of fertility and mechanical composition are equipped with practically the same machinery and farming methodology.

1. Systematization of parameters for assessing the effectiveness of land use regulation

The effectiveness of land resources use should have quantitative parameters characterizing each of the aspects directly involved in the process of land use regulation.

The integrated performance indicator contains a generalized assessment of the phenomenon, and assumes the existence of a plurality of indicators that quantitatively reflect different aspects of the processes taking place. So, for example, the criterion of economic efficiency reflects the generalized result of an economic phenomenon on the basis of which evaluation, definition and classification of partial indicators of efficiency are carried out. It is associated with the use of limited resources and the production of consumer goods. Many scientists researching the field of land use divide efficiency into several types: economic and ecological. This indicates the expediency of recognizing the multivariate performance criterion.

The concept of effective land use is very complex and multifaceted and can not be clearly defined. Highlighted in one of its aspects, it can be formulated incorrectly in the sense of the complex meaning of the concept itself. Essence effective land use in agricultural production is formed by two main aspects: economic and ecological.

To define the economic efficiency of land resources use, it is expedient to apply quantitative resource parameters: an indicator of the efficiency of capital investments, resource potential, intensification of agricultural production, etc. At the same time, each of the listed parameters is a complex feature that includes components elaborating it. In particular, the concept of intensification of land resources use in a market economy is inextricably linked with the territorial organization of new forms of management, which, takes into account natural and landscape conditions, as well as forms of ownership established in modern land relations.

Environmental component is important from the point of view of efficiency assessment [7], which in land use is characterized by three main groups of parameters: soil quality criteria, natural landscape conditions, as well as pollution and degradation of land resources.

Based on the analysis and parameters characterizing the economic and ecological components of land use, a set of indicators was formed that will determine the effectiveness of land use development.

A set of parameters for each of the environments (ecological, economic) determines indicators that will be used to assess the effectiveness of land use regulation.

The main indicators in terms of the economic environment include:

- agricultural production growth index;
- agricultural products per capita;
- the amount of capital investment per hectare of agricultural land;
- average monthly nominal wage of full-time employees in agriculture field.

Indicators characterizing the ecological environment include:

- share of agricultural arable land under organic fertilization;
- level of plowing of agricultural lands;
- share of eroded land;
- weighted average humus content.

2. Aggregated indicators

The next step will be defining of aggregated indicators, that is, indicators that are calculated from several separate basic indicators and characterize the state of related group indicators of the economic and ecological environment. We will determine them by the arithmetic mean method (table 1).

Each basic indicator is evaluated on a four-point scale. The highest score is assigned to the indicator that is closest to the benchmark value. The maximum number of points per group is 16.

The integrated indicator is calculated according to the formula of the arithmetic mean among the defined aggregated indicators.

Unlike ecological indicators [1], economic basic indicators belong mainly to positive indicators, the growth of which enhance the development of the ecological and economic system to the reference state.

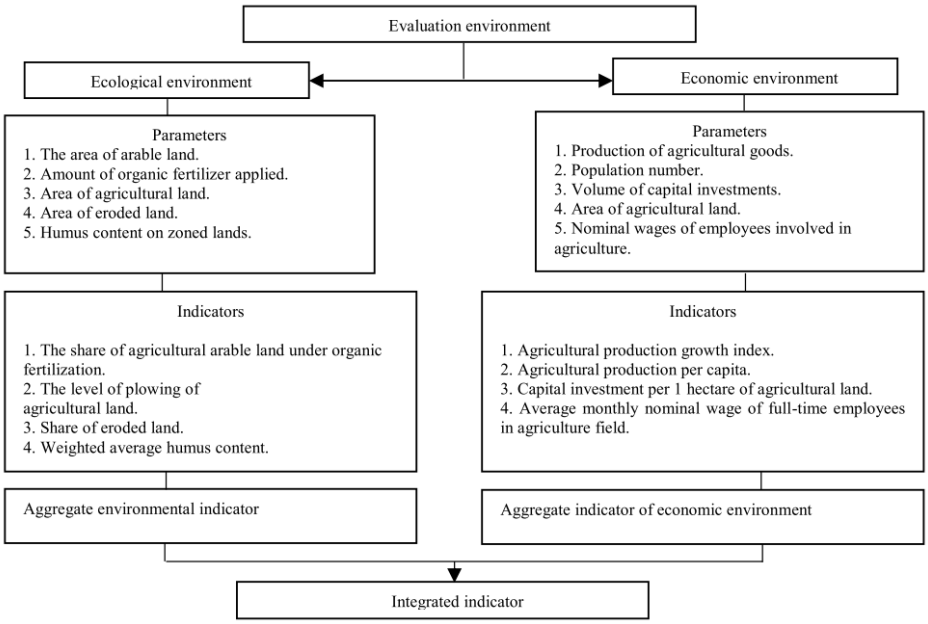


Fig. 1. Systematization of parameters for assessing the effectiveness of land use regulation

Table 1. Calculation of aggregated indicators

No	Aggregate indicator	Basic indicators	Calculation of the aggregate indicator
1	Economic environment	Agricultural production growth index (X_1)	$AIecon. = \frac{X_1 + X_2 + X_3 + X_4}{16}$
		Agricultural production per capita (X_2)	
		The volume of capital investments per 1 ha of agricultural land (X_3)	
		Average monthly nominal wage of full-time employees in agriculture field (X_4)	
2	Ecological environment	Share of agricultural arable land under organic fertilization (U_1)	$AIecol. = \frac{U_1 + U_2 + U_3 + U_4}{16}$
		The level of plowing of agricultural lands (U_2)	
		Share of eroded land (U_3)	
		Weighted average humus content (U_4)	

Table 2. Basic indicators of land use efficiency assessment

No	Group	Indicators	Formula	Conditional notation of parameters
1	2	3	4	5
1	Economic indicators	Agricultural production growth index, %	$Ap = \frac{P\ r.y.}{P\ p.y.} - 100\%$	where *P r.y.* – is agricultural production of the reporting year, in million hryvnias, *P p.y.* – is agricultural production of the previous year, million hryvnias
		Agricultural production per capita, thousand hryvnias	$P_1 = \frac{P\ r.y.}{pop} * 1000$	where *pop* – is the population, million people
		Capital investment per 1 hectare of agricultural land, thousand hryvnias/ha	$KI_{1ha} = \frac{KI_{tot}}{P_a} * 1000$	where KI_{tot} – is amount of capital investments in agriculture, in million hryvnias, P_a – is area of agricultural land, million ha
		Average monthly nominal wage of full-time employees in agriculture field, thousand hryvnias	$P_m = \frac{W}{Q}$	where W – is the wage fund for full-time employees in agriculture, thousand UAH, Q – is the number of full-time employees in agriculture, people.
2	Environmental indicators	Share of agricultural arable land under organic fertilization, %	$O = \frac{A_O}{A_T} * 100\%$	A_O – is the area of agricultural land to which organic fertilizers were applied, thousand hectares, A_T – is the total area of agricultural land, thousand hectares
		The level of plowing of agricultural lands, %	$L_P = \frac{A_T}{A_a} * 100\%$	A_T – is the area of agricultural land, thousand ha, A_a – is the area of tillage
		Share of eroded land, %	$S = \frac{A_E}{A_T} * 100\%$	A_E – is the area of eroded lands, thousand ha
		Weighted average humus content, %	$G_A = \frac{\sum G_i}{N} * 100\%$	$\sum G_i$– is the sum of humus contents by agricultural zones, %, N – is a number of land units according to zonal agricultural distribution

The integrated indicator is calculated according to the formula of the arithmetic mean among the defined aggregated indicators. The basic indicators and the parameters that form them are presented in more detail in table 2.

The developed systematization of the parameters of land use regulation indicates their multifactorial nature, interdependence, multidirectionality in optimization issues, which may involve the development of additional procedures for making management decisions in practical use.

When refining the system of parameters for assessing the effectiveness of land use regulation, we proceeded from the following:

– the achieved level of land use efficiency at a given period of time and the necessary environmental and economic prerequisites create the potential for its improvement;
– the economic efficiency of land use should reflect the real results used in the management process;
– the economic efficiency of land resources use is an integral part of the intensification of agricultural production, including the agricultural reproduction process.

Thus, the aggregate concept of the land use regulation efficiency (*I*) can be analytically presented as a functional dependence of economic (*E*) and environmental indicators (*EI*):

$$I = f(E, EI) \quad (1)$$

The practical application of this functionality to solve the problem of increasing the efficiency of land use can be implemented in the mechanism of state regulation of land use, one of the stages of which is the systematization of ideas on modern directions of increasing the efficiency of land use.

The calculation of the aggregated and integral indicator of the efficiency of land use regulation is in the table 3.

Table 3. Determination of aggregate and integral indicators of the effectiveness of land use regulation in the Rivne region in Ukraine in 2021

No	Indicators	Scores / Values				Actual value	Point	Aggregate indicator
		1	2	3	4			
Economic indicators								
1	Agricultural production growth index, %	<5	5–10	10–20	> 20	2	1	0.438
2	Agricultural production per capita, thousand hryvnias	<10	10–20	20–30	> 30	15.4	2	
3	Capital investment per 1 hectare of agricultural land, thousand hryvnias/ha	0–1	1–2	2–3	> 3	1,2	2	
4	Average monthly nominal wage of full-time employees in agriculture field, thousand hryvnias	0–7	7–10	10–13	> 13	9.5	2	
Environmental indicators								
1	Share of agricultural arable land under organic fertilization, %	0–2	2–5	5–10	>10	11	4	0.813
2	The level of plowing of agricultural land, %	>55	55–45	45–35	<35	33	4	
3	Share of eroded land, %	>60	60–40	40–20	20–0	34.6	3	
4	Weighted average humus content, %	1–2	2–3	3–4	> 4	2,3	2	
Integral indicator of efficiency I = 0.626 (or 62.6%)								

3. Conclusion

The article defines aggregate and integral indicators of the effectiveness of land use regulation in the Rivne region in Ukraine. The aggregate indicators in terms of economic and environmental components were 0.438 and 0.813, respectively. The aggregate environmental indicator is almost 2 times higher than the aggregated economic one, which indicates a better state of the environmental component than the economic one. This was due to the maximum score for such indicators as the share of agricultural arable land under organic fertilizers and the level of plowed agricultural land. The integral indicator was calculated as the arithmetic mean of the aggregated indicators and it is 0.626 or 62.6%.

References

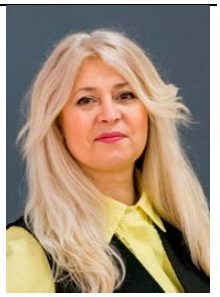
[1] Borel-Saladi J. M., Turo I. N.: The green economy: incremental change or transformation? Environ. Policy Gov 23(4), 2013, 209–220.
[2] Dobriak D. S.: Economic turnover of land in Ukraine: theory, methodology and practice. Harvest, Kiev 2004.
[3] Harnaha O. M.: Basics of effective use of land resources: monograph. NUWEE, Rivne 2015.
[4] Harnaha O. M.: Organizational and economic measures to restore soil fertility. Bulletin of the National University of Water and Environmental Engineering: Collection of scientific papers. Economy 4(36), 2006, 41–46.
[5] Hutsuliak H.: Land planning of rural areas and the formation of optimal structures land use. Land manager newsletter 3, 2013, 46–48.
[6] Kalimo H., Lifset R., van Rossem C., van Wassenhove L.: Greening the economy through design incentives: Allocating extended producer responsibility. Eur. Energy & Envtl. L. Rev. 21, 2012, 274.
[7] Kenis A., Lievens M.: Greening the economy or economizing the green project? When environmental concerns are turned into a means to save the market. Rev Radic Polit Econ 48(2), 2016, 217–234.
[8] Rasmussen M. B.: Greening the Economy: Articulation and the Problem of Governance in the Andes. MRD 32(2), 2012, 149–157.
[9] Sokhnych A. Ya.: Ecologization of land use. Land Management Newsletter 2, 2005, 19–23.
[10] Tretiak A. M.: Ecology of land use: theoretical and methodological foundations of formation and administration: monograph. Gryn, Kherson 2012.

Ph.D. Oleksandr Harnaha
e-mail: o.m.gharnagha@nuwm.edu.ua

Associate professor of the Institute of Economics and Management, National University of Water and Environmental Engineering, Ukraine. Author and co-author of more than 60 scientific papers. Scientific interests are based on the study of the nature use economics, the development of the land market etc.

http://orcid.org/0000-0002-5236-7299

D.Sc. Natalia Savina
e-mail: n.b.savina@nuwm.edu.ua



Doctor of Economics of the Institute of Economics and Management, National University of Water and Environmental Engineering, Ukraine. Author and co-author of more than 100 scientific papers. The author`s research area focuses on the study of economics and management of the national economy.

http://orcid.org/0000-0001-8339-1219

M.Sc. Volodymyr Hrytsiuk
e-mail: v.v.hrytsiuk@nuwm.edu.ua

Senior lecture of the Nadsluchansky Institute, National University of Water and Environmental Engineering, Ukraine.
Scientific interests include the study of nature management aspects, economics of forestry and agriculture etc.

http://orcid.org/0009-0000-2270-0874

LIST OF REVIEWERS COOPERATING WITH THE JOURNAL IN 2023

LISTA RECENZENTÓW WSPÓŁPRACUJĄCYCH Z CZASOPISMEM W ROKU 2023

Prof.	Avrunin	Oleg	Kharkiv National University of Radio Electronics (Kharkiv, Ukraine)
Prof.	Benba	Achraf	Mohammed V University in Rabat, National School of Arts and Crafts (Rabat, Marocco)
Prof.	Bezuglyi	Mykhailo	National Technical University of Ukraine „Kyiv Polytechnic Institute" (Kyiv, Ukraine)
Ph.D.	Bober	Dariusz	University of Rzeszow (Rzeszow, Poland)
Ph.D.	Boryczko	Krzysztof	University of Rzeszow (Rzeszow, Poland)
Prof.	Boyko	Oksana	Danylo Halytsky Lviv National Medical University (Lviv, Ukraine)
Ph.D.	Buczaj	Marcin	Lublin University of Technology (Lublin, Poland)
Prof.	Chodakowska	Ewa	Bialystok University of Technology (Bialystok, Poland)
Ph.D.	Chodorek	Agnieszka	Kielce University of Technology (Kielce, Poland)
Prof.	Dembitska	Sofia	Vinnytsia National Technical University (Vinnytsia, Ukraine)
Prof.	Duraj	Agnieszka	Lodz University of Technology (Lodz, Poland)
Ph.D.	Dzieńkowski	Mariusz	Lublin University of Technology (Lublin, Poland)
Ph.D.	Dzierżak	Róża	Lublin University of Technology (Lublin, Poland)
Ph.D.	Fedotova	Natali	Sumy State University (Sumy, Ukraine)
Prof.	Gromaszek	Konrad	Lublin University of Technology (Lublin, Poland)
Prof.	Hotra	Oleksandra	Lublin University of Technology (Lublin, Poland)
Prof.	Junisbekov	Mukhtar	M. Kh. Dulaty Taraz State University (Taraz, Kazakhstan)
Ph.D.	Kalizhanova	Aliya	Institute of Information and Computational Technologies SR MES RK (Almaty, Kazakhstan), Al-Farabi Kazakh, National University (Almaty, Kazakhstan)
Ph.D.	Kłosowski	Grzegorz	Lublin University of Technology (Lublin, Poland)
Ph.D.	Kociubiński	Andrzej	Lublin University of Technology (Lublin, Poland)
Prof.	Kołtunowicz	Tomasz	Lublin University of Technology (Lublin, Poland)
Prof.	Kotyra	Andrzej	Lublin University of Technology (Lublin, Poland)
Ph.D.	Kovalchuk	Nadiia	Lutsk National Technical University (Lutsk, Ukraine)
Ph.D.	Kwedlo	Wojciech	Bialystok University of Technology (Bialystok, Poland)
Ph.D.	Lach	Zbigniew	Lublin University of Technology (Lublin, Poland)
Ph.D.	Łukasik	Edyta	Lublin University of Technology (Lublin, Poland)
Prof.	Lytvynenko	Volodymyr	Kherson National Technical University (Kherson, Ukraine)
Ph.D.	Maciejewski	Marcin	Lublin University of Technology (Lublin, Poland)
Ph.D.	Mamyrbayev	Orken	Institute of Information and Computational Technologies SR MES RK (Almaty, Kazakhstan)
Prof.	Mikulka	Jan	Brno University of Technology (Brno, Czech Republic)
Ph.D.	Naumchuk	Oleksandr	National University of Water and Environmental Engineering (Rivne, Ukraine)
Prof.	Nazarenko	Ivan	Kyiv National University of Construction and Architecture (Kyiv, Ukraine)
Prof.	Nykyforova	Larysa	National University of Life and Environmental Sciences of Ukraine (Kyiv, Ukraine)
Prof.	Omiotek	Zbigniew	Lublin University of Technology (Lublin, Poland)
Prof.	Pastuszak	Grzegorz	Warsaw University of Technology (Warsaw, Poland)
Prof.	Pavlov	Sergey	Vinnytsia National Technical University (Vinnytsia, Ukraine)
Ph.D.	Pawłowski	Eligiusz	Lublin University of Technology (Lublin, Poland)
Ph.D.	Plechawska-Wójcik	Małgorzata	Lublin University of Technology (Lublin, Poland)
Prof.	Polishchuk	Leonid	Vinnytsia National Technical University (Vinnytsia, Ukraine)
Prof.	Prokopovych	Ihor	Odessa National Polytechnic University (Odessa, Ukraine)
Prof.	Pryshchepa	Alla	National University of Water and Environmental Engineering (Rivne, Ukraine)
Prof.	Savina	Natalia	National University of Water and Environmental Engineering (Rivne, Ukraine)
Prof.	Semenov	Andriy	Vinnytsia National Technical University (Vinnytsia, Ukraine)
Prof.	Shchur	Taras	Cyclone Manufacturing Inc (Ontario, Canada)
Prof.	Sikora	Jan	Research & Development Centre Netrix S.A. (Lublin, Poland), WSEI University (Lublin, Poland)
Ph.D.	Smailova	Saule	D. Serikbayev East Kazakhstan State Technical University (Ust-Kamenogorsk, Kazakhstan)
Prof.	Smolarz	Andrzej	Lublin University of Technology (Lublin, Poland)
Ph.D.	Stakhova	Anzhelika	Slovak University of Technology (Bratislava, Slovakia)
Ph.D.	Stelmakh	Nataliia	National Technical University of Ukraine „Kyiv Polytechnic Institute" (Kyiv, Ukraine)
Prof.	Sulemenov	Batyrbek	Kazakh National Research Technical University after K.I.Satpayev (Almaty, Kazakhstan)
Ph.D.	Surtel	Wojciech	Lublin University of Technology (Lublin, Poland)
Ph.D.	Susik	Robert	Lodz University of Technology (Lodz, Poland)
Prof.	Szromba	Andrzej	Cracow University of Technology (Cracow, Poland)
Prof.	Titova	Nataliia	Odessa National Polytechnic University (Odessa, Ukraine)
Prof.	Ushenko	Alexander G.	Yuriy Fedkovych Chernivtsi National University (Chernivtsi, Ukraine)
Prof.	Vasilevskyi	Oleksandr	University of Texas at Austin (Austin, USA)
Ph.D.	Warda	Piotr	Lublin University of Technology (Lublin, Poland)
Prof.	Yashchyshyn	Yevhen	Warsaw University of Technology (Warsaw, Poland)
Ph.D.	Yesmakhanova	Laura	M. Kh. Dulaty Taraz State University (Taraz, Kazakhstan)
Prof.	Zabolotnii	Serhii	Cherkasy State Business-College (Cherkasy, Ukraine)
Ph.D.	Zagorodna	Nataliya	Ternopil Ivan Puluj National Technical University (Ternopil, Ukraine)
Prof.	Zdunek	Rafał	Wroclaw University of Science and Technology (Wroclaw, Poland)