МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧЕРКАСЬКИЙ ДЕРЖАВНИЙ ТЕХНОЛОГІЧНИЙ УНІВЕРСИТЕТ

ВІСНИК
ЧЕРКАСЬКОГО ДЕРЖАВНОГО ТЕХНОЛОГІЧНОГО УНІВЕРСИТЕТУ

Науковий збірник

Том 31,
№ 2. 2026

ЧЕРКАСИ
2026

ISSN: 2306-4412
E-ISSN: 2708-6070

Засновник:
Черкаський державний технологічний університет

Рік заснування: 1996
Рекомендовано до друку та поширення
через мережу Інтернет Вченою радою
Черкаського державного технологічного університету
(протокол № 13 від 18 травня 2026 р.)

Державна реєстрація: Ідентифікатор медіа R30-04613.
Рішення Національної ради України з питань телебачення і радіомовлення
№ 1916, протокол № 17 (30.05.2024).

Науковий збірник входить до переліку наукових фахових видань України
Категорія «Б». Спеціальність:
технічні фізико-математичні (Наказ Міністерства освіти і науки України № 886 від 02.07.2020):
113 Прикладна математика, 121 Інженерія програмного забезпечення, 122 Комп'ютерні науки, 123 Комп'ютерна інженерія, 125 Кібербезпека та захист інформації,
126 Інформаційні системи та технології, 131 Прикладна механіка, 132 Матеріалознавство,
133 Галузеве машинобудування, 151 Автоматизація та комп'ютерно-інтегровані технології,
152 Метрологія та інформаційно-вимірювальна техніка,
172 Телекомунікації та радіотехніка; технічні (Наказ Міністерства освіти і науки України № 1188 від 24.09.2020): 101 Екологія, 161 Хімічні технології та інженерія

Науковий збірник представлено у міжнародних наукометричних базах даних, репозитаріях та пошукових системах: Bielefeld Academic Search Engine (BASE), Crossref, CiteFactor, Litmaps, Ulrich's Periodicals Directory, WorldCat,
J-Gate, Open Ukrainian Citation Index (OUCI),
Наукова періодика України, Національна бібліотека України імені В. І. Вернадського (НБУВ),
Dimensions, UCSB Library, Google Академія, German Union Catalogue of Serials (ZDB),
Бібліотека Університету Осло, Бібліотека Університету Халла,
SOLO - Search Oxford Libraries Online, Інститут Європейського університету, Бібліотека Лейпцизького університету (UBL), Бібліотека Кембриджського університету

Адреса редакції:
Черкаський державний технологічний університет
18006, бульв. Шевченка, 460, м. Черкаси, Україна
E-mail: info@bulletin-chstu.com.ua
https://bulletin-chstu.com.ua/uk

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
CHERKASY STATE TECHNOLOGICAL UNIVERSITY

BULLETIN
OF CHERKASY STATE TECHNOLOGICAL UNIVERSITY

Scientific Collection

Volume 31,
No. 2. 2026

CHERKASY
2026

ISSN: 2306-4412
E-ISSN: 2708-6070

Founder:
Cherkasy State Technological University

Year of foundation: 1996
Recommended for printing and distribution
via the Internet by the Academic Council
of Cherkasy State Technological University
(Minutes No. 13 of May 18, 2026)

State Registration: Media identifier R30-04613.
Decision of the National Council of Television and Radio Broadcasting of Ukraine
No. 1916, Minutes No. 17, dated 30.05.2024.

The scientific collection is included in the list of Professional Scientific Publications of Ukraine
Category "B". Specialties: technical, physical and mathematical (Order of the Ministry of Education and Science of Ukraine No. 886 of 02.07.2020): 0541 Mathematics,
0612 Database and network design and administration,
0613 Software and applications development and analysis, 0714 Electronics and automation,
0715 Mechanics and metal trades, 0716 Motor vehicles, ships and aircraft,
0788 Inter-disciplinary programmes and qualifications involving engineering, manufacturing and construction; technical (Order of the Ministry of Education and Science of Ukraine No. 1188 of 24.09.2020): 0521 Environmental sciences, 0522 Natural environments and wildlife,
0711 Chemical engineering and processes

The scientific collection is presented in the international scientometric databases, repositories and scientific systems: Bielefeld Academic Search Engine (BASE), Crossref, CiteFactor, Litmaps,
Ulrich's Periodicals Directory, WorldCat, J-Gate,
Open Ukrainian Citation Index (OUCI),
Scientific Periodicals of Ukraine, Vernadsky National Library of Ukraine (VNLU), Dimensions,
UCSB Library, Google Scholar, German Union Catalogue of Serials (ZDB), University of Oslo Library,
University of Hull Library, SOLO - Search Oxford Libraries Online, European University Institute,
Leipzig University Library (UBL), Cambridge University Library

Editors office address:
Cherkasy State Technological University
18006, 460 Shevchenko Blvd., Cherkasy, Ukraine
E-mail: info@bulletin-chstu.com.ua
https://bulletin-chstu.com.ua/en

Редакційна колегія

Головний редактор:

Еміль Фауре	Доктор технічних наук, професор кафедри інформаційної безпеки та комп'ютерної інженерії, проректор з науково-дослідної роботи та міжнародних зв'язків, Черкаський державний технологічний університет, Україна

Заступник головного редактора

Джаміль Аль-Аззех	Доктор філософії у галузі комп'ютерної інженерії, професор, професор кафедри комп'ютерної інженерії, прикладний університет Аль-Балка, Йорданія

Національні члени редколегії:

Володимир Артемчук	Доктор технічних наук, старший науковий співробітник, заступник директора з науково-організаційної роботи, Інститут проблем моделювання в енергетиці ім. Г. Є. Пухова Національної академії наук України, Україна
Юлія Бондаренко	Кандидат технічних наук, професор, старший науковий співробітник Державного науково-дослідного інституту випробувань і сертифікації озброєння та військової техніки, Україна
Вячеслав Ващенко	Доктор технічних наук, професор, завідувач кафедри фундаментальних дисциплін та прикладного матеріалознавства, Черкаський державний технологічний університет, Україна
Сергій Заболотній	Доктор технічних наук, професор, професор кафедри комп'ютерної інженерії та інформаційних технологій, «Черкаський державний бізнес-коледж», Україна
Володимир Палагін	Доктор технічних наук, професор, завідувач кафедри робототехнічних і телекомунікаційних систем та кібербезпеки, Черкаський державний технологічний університет, Україна
Євген Федоров	Доктор технічних наук, професор, професор кафедри робототехніки та спеціалізованих комп'ютерних систем, Черкаський державний технологічний університет, Україна
Костянтин Базіло	Доктор технічних наук, професор, професор кафедри приладобудування, мехатроніки та комп'ютеризованих технологій, Черкаський державний технологічний університет, Україна
Наталія Бойко	Кандидат економічних наук, доцент, доцент кафедри систем штучного інтелекту, Інститут комп'ютерних наук та інформаційних технологій, Національний університет «Львівська політехніка», Україна
Тетяна Нескородєва	Доктор технічних наук, професор, професор кафедри інформаційних технологій, Уманський національний університет, Україна
Максим Бондаренко	Доктор технічних наук, професор, завідувач кафедри приладобудування, мехатроніки та комп'ютеризованих технологій, Черкаський державний технологічний університет, Україна

Міжнародні члени редколегії:

Максим Явіч	Доктор філософії, професор, Кавказький університет, Грузія
Томас Ганне	Кандидат економічних наук, професор, професор інформаційних систем, Університет прикладних наук і мистецтв Північно-Західної Швейцарії, Швейцарія

Міжнародні члени редколегії:

Дженґіз Хакан Айдін	Доктор філософії у галузі відкритої освіти, професор, Анатолійський університет, Туреччина
Вітор А. Кунья	Доктор філософії з комп'ютерної інженерії, дослідник Інституту телекомунікацій, Університет Авейру, Португалія
Сір'є Віркус	Доктор філософії у галузі інформаційних та комунікаційних досліджень, професор, професор інформаційних наук Школи цифрових технологій,Талліннський університет, Естонія
Теро Вартяйнен	Доктор філософії у галузі інформаційних систем, професор, професор факультету технологій та інновацій, Університет Вааса, Фінляндія

Editorial Board

Editor-in-Chief:

Emil Faure — Doctor of Technical Sciences, Professor of the Department of Information Security and Computer Engineering, Vice-Rector for Research and International Relations, Cherkasy State Technological University, Ukraine

Deputy Editor-in-Chief

Jamil Al Azzeh — PhD in Computer Engineering, Professor, Professor of the Computer Engineering Department, Al-Balqa' Applied University, Jordan

National Members of the Editorial Board:

Volodymyr Artemchuk — Doctor of Technical Sciences, Senior Researcher, Deputy Director for Scientific and Organisational Work, G.E. Pukhov Institute for Modelling in Energy Engineering of National Academy of Sciences of Ukraine, Ukraine

Iuliia Bondarenko — PhD in Technical Sciences, Professor, Senior Researcher at the State Research Institute for Testing and Certification of Arms and Military Equipment, Ukraine

Viacheslav Vashchenko — Doctor of Technical Sciences, Professor, Head of the Department of Fundamental Disciplines and Applied Materials Science, Cherkasy State Technological University, Ukraine

Serhii Zabolotnii — Doctor of Technical Sciences, Professor, Professor of the Department of Computer Engineering and Information Technologies, Cherkasy State Technological University, Ukraine

Volodymyr Palahin — Doctor of Technical Sciences, Professor, Head of the Department of Robotics, Telecommunications Systems and Cybersecurity, Cherkasy State Technological University, Ukraine

Eugene Fedorov — Doctor of Technical Sciences, Professor, Professor of the Department of Statistics and Applied Mathematics, Cherkasy State Technological University, Ukraine

Kostiantyn Bazilo — Doctor of Technical Sciences, Professor, Professor of the Department of Instrumentation, Mechatronics and Computerised Technologies, Cherkasy State Technological University, Ukraine

Nataliya Boyko — PhD in Economics, Associate Professor, Associate Professor at the Department of Artificial Intelligence Systems, Institute of Computer Science and Information Technologies, Lviv Polytechnic National University, Ukraine

Tetyana Neskorodyeva — Doctor of Technical Sciences, Professor, Professor of the Department of Information Technologies, Uman National University, Ukraine

Maksym Bondarenko — Doctor of Technical Sciences, Professor, Head of the Department of Instrumentation, Mechatronics and Computerised Technologies, Cherkasy State Technological University, Ukraine

International Members of the Editorial Board:

Maksim Iavich — Doctor of Philosophy, Professor, Caucasus University, Georgia

Thomas Hanne — PhD in Economics, Professor, Professor of Information Systems, University of Applied Sciences and Arts Northwestern Switzerland, Switzerland

International Members of the Editorial Board:

Cengiz Hakan Aydın	PhD in Open Education, Professor, Director of the Office of Teaching and Learning Enhancement, Özyeğin University, Turkey
Vitor A. Cunha	PhD, Researcher at the Institute of Telecommunications, University of Aveiro, Portugal
Sirje Virkus	PhD in Information and Communication Studies, Professor, Professor of Information Science,Tallinn University, Estonia
Tero Vartiainen	PhD in Information Systems, Professor, Professor at the Faculty of Technology and Innovation, University of Vaasa, Finland

Зміст / Contents

BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 2. 2026

UDC 004.932.2:004.056
DOI: 10.62660/bcstu/2.2026.11

ISSN: 2306-4412
E-ISSN: 2708-6070

Method for detecting the results of digital image cloning in post-processing

Svitlana Hryhorenko*
PhD in Technical Sciences
National University "Odesa Polytechnic"
65044, 1 Shevchenko Ave., Odesa, Ukraine
https://orcid.org/0009-0006-4551-8243

Oleksandr Yevseiev
Senior Lecturer
National University "Odesa Polytechnic"
65044, 1 Shevchenko Ave., Odesa, Ukraine
https://orcid.org/0009-0009-3100-2643

Mykola Voronov
Postgraduate Student
National University "Odesa Polytechnic"
65044, 1 Shevchenko Ave., Odesa, Ukraine
https://orcid.org/0009-0003-2816-0495

Abstract. The relevance of this study is driven by the increasing number of digital image falsification cases and the widespread use of fragment cloning as a common manipulation technique, which poses serious challenges to information security. The aim of this research was to develop and experimentally validate a robust method for detecting digital image cloning under post-processing conditions based on the analysis of the minimum mean block difference matrix, enabling reliable identification and spatial localisation of cloned regions and their original prototypes. The proposed method was based on iterative interval bisection combined with binary cross-analysis of block differences and was implemented in the Matlab environment. The experimental framework included modelling additive Gaussian, multiplicative and impulse noise, JPEG compression, spatial filtering, and combined distortions. Visual fidelity was quantitatively evaluated using the peak signal-to-noise ratio with a threshold value of signal-to-noise ratio PSNR > 37 dB. Computational experiments were conducted on central regions of digital images with a resolution of 512 × 512 pixels. Cloned regions occupied less than 0.85% of the image area, typically ranging from 0.098% to 0.39%, while smaller cloned areas below 0.098% were also considered. The analysis was performed on a randomly selected single colour channel using block-based processing. In the absence of post-processing distortions, the global minimum of the minimum mean block difference function reached zero, resulting in maximum detection performance with a true positive rate of 100%. False positive rate evaluation on original images yielded values of 11%, 7.5%, and 3% for block sizes of 16, 24, and 32, respectively. Although the proposed approach demonstrates high detection sensitivity, the results indicate limitations in suppressing false positives compared to some state-of-the-art methods, highlighting the need for further optimisation

Keywords: peak signal-to-noise ratio; "salt and pepper"; gaussian noise; multiplicative noise; "intersection"; Matlab

Article's History: Received: 11.12.2025; Revised: 20.03.2026; Accepted: 18.05.2026; Published: 26.06.2026

Suggested Citation:
Hryhorenko, S., Yevseiev, O., & Voronov, M., (2026). Method for detecting the results of digital image cloning in post-processing. *Bulletin of Cherkasy State Technological University*, 31(2), 11-26. doi: 10.62660/bcstu/2.2026.11.

*Corresponding author (s.m.hryhorenko@op.edu.ua)

Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/)

INTRODUCTION

The rapid development of digital technologies and the widespread use of digital images (DI) in the media, social networks, electronic document management systems, digital forensics and information security are leading to an increase in the risks associated with the falsification of visual information. One of the most common and, at the same time, most difficult methods of forgery to detect automatically is the cloning of fragments of digital images, which allows objects to be hidden or duplicated without the use of external image sources. The task of detecting the results of cloning becomes particularly difficult in the context of post-processing of digital images, which almost always accompanies the process of creating or distributing images. Noise overlay, filtering, lossy compression, changing storage formats, and other disruptive influences significantly complicate the detection of cloning traces, as they mask the characteristic features of forgery and reduce the similarity between the corresponding fragments of the clone and its prototype. In real-world scenarios of falsification, such influences are not an exception but a typical practice that requires the development of specialised analysis methods that are resistant to post-processing.

Thus, in the work of G. Gani & F. Qadir (2020), a block-oriented approach was proposed in which discrete cosine transform (DCT) is used to form compact frequency features that are less sensitive to noise, JPEG artefacts, and blurring. Additionally, the use of cellular automata allows for the effective grouping of similar blocks and reduces the number of false positives after post-processing, which ensures increased stability of the matching procedure in combined attacks. A.U. Shehin & D. Sankar (2024) used DCT features together with eigenvalues to form an invariant representation of blocks that preserves discriminatory properties after JPEG compression and blur filtering. The proposed scheme also provides robustness to rotations, allowing stable localisation of forged areas under conditions of complex post-processing and geometric transformations. P. Duszejko *et al.* (2025) the scientific literature and research findings, focusing on techniques that detect image manipulations and localise areas of tampering, utilising both statistical properties of images and embedded hidden watermarks. Passive methods, based on analysing the image itself, are versatile and can be applied across a broad range of cases; however, their effectiveness depends on the complexity of the modifications and the characteristics of the image. Active methods, which involve embedding additional information into the image, offer precise detection and localisation of changes but require complete control over creating and distributing visual materials. Both approaches have their applications depending on the context and available resources.

The work F. Li *et al.* (2022). recommends using Graph Neural Networks, Convolutional Neural Networks, Visual Geometry Group Network with 16 Layers, MobileNet, and Residual Network with 50 Layers. Method uniquely integrates these architectures to effectively detect and analyse multiple types of image forgeries, including image splicing and copy-move forgeries. A. Akram *et al.* (2025) argued that the growing reliance on digital image processing technologies is partly explained by the widespread availability of reliable and efficient image capture technologies. The ease with which image content can be altered using modern image editing technologies creates new challenges for forensic analysis. A structured hybrid system for searching for important objects in images is presented.

Despite the existence of a significant number of approaches to detecting cloning in digital images, most of them demonstrate limited effectiveness in the case of combined disruptive influences or require significant computational resources, complex parameter tuning, or prior training on large data sets. In this context, it is relevant to develop a method for detecting the results of DI cloning in post-processing conditions that combines mathematical soundness, algorithmic transparency, and practical applicability for use in expert and automated analysis systems. The purpose of the article was to develop and experimentally substantiate a method for detecting DI cloning results in post-processing conditions, which is based on the analysis of the minimum block difference matrix and allows for the stable detection and localisation of a clone and its prototype in the presence of noise, lossy compression, and other interfering influences.

MATERIALS AND METHODS

In order to make an informed choice of disturbance parameters that ensure the reliability of the visual perception of digital images, a computational experiment was conducted in the Matlab environment. Computational studies were performed using licensed Matlab software (Kadam *et al.*, 2022; Ghai *et al.*, 2024). The Matlab platform was selected due to its extensive set of built-in functions for image processing, numerical computation, and data visualisation, which enables the implementation of complex experimental pipelines and ensures the reproducibility of results.

For simplicity of presentation, without limiting the generality of the considerations, monochrome DI are considered at this stage of the study. Any DI is the result of discretisation of the domain $D \subset R^2$ and quantisation of the function values $f(x, y)$, $(x, y) \in D$, resulting in $n \times m$ – matrix F with elements $f_{ij} \in \{0,1,\ldots, 255\}$, $i = 1, n$, $J = 1, m$ and the geometric representation can be a three-dimensional surface S, which is the graph of the function $g(x, y)$, approximating $f(x, y)$ on a multiplicity D of vicoristan elements F, in particular $g(x, y)$ can act as an interpolation spline for elements F, which is assumed below. It is shown $g(x, y)$ that if is a 1st degree interpolation spline, then the task of its construction is insensitive to changes in the input data – the brightness values of the pixels

of the image. The defined parameter characterising the surface S and, as a result DI, the matrix F as a whole, and at the same time insensitive to small changes in the characteristics of the surface (DI) – V the volume of a cylindrical body T with a generatrix parallel to the axis of the applicator OZ, limited by the surface S and the coordinate plane of the characteristics of differences: $V = \iint g(x, y)\, dxdy$, $g(x, y)$ – 1st degree interpolation spline for elements F.

If $K \subset D$, $P \subset D$ – areas corresponding to areas of the clone/prototype (CP), then they can be matched: $V_k = \iint g(x, y)\, dxdy$, $V_p = \iint g(x, y)\, dxdy$. If the central point was not post-processed, then: $V_k = V_p$, otherwise, due to the insensitivity of the volume calculation problem: $V_k \approx V_p$.

For numerical calculation of volume values using double integrals as a composite quadrature formula, the integral sum for a double integral is used, corresponding to the division of the integration region $L \subset D$, for which the partial region is a square, in the centre of which there is a point with the coordinates of the corresponding pixel. Taking into account that the coordinates of adjacent pixels differ by one unit, it can be assumed that the length of the side of each resulting square, and therefore its area S_{ij}, is equal to 1.

$$\iint_L g(x,y)\,dxdy = \sum_{(i,j)\in L} g(i,j)S_{ij} = \sum_{(i,j)\in L} g(i,j) = \sum_{(i,j)\in L} f_{ij}. \quad (1)$$

For (1), the degree of the interpolation spline does not matter , which eliminates the need to consider a spline of degree higher than one for interpolating the elements of the matrix F in the course of the proposed solution of the problem considered in the work.

Thus, for areas of the CP without subsequent processing of the DI, the following relationship must be satisfied, for which insensitivity to disturbances in the input data is established:

$$\sum_{(i,j)\in P} f_{ij} = \sum_{(i,j)\in K} f_{ij}. \quad (2)$$

And under conditions of additional disruptive actions (DA):

$$\sum_{(i,j)\in K} f_{ij} \approx \sum_{(i,j)\in P} f_{ij}. \quad (3)$$

It is assumed that post-processing after cloning applies to the entire DI. Taking into account the results of the literature review, a block-oriented approach is used for the developed method as the most accurate among the existing ones. It is further assumed that the areas of the CP regions are such that the prototype has at least one $q \times q$ – block that does not belong to the CP intersection area.

Taking into account the experiment, during which the presence $q = 4$, $q = 8$ and $q \geq 16$ the absence of blocks in the original DI $q \times q$ – block for which condition (2) was satisfied were established, the criterion for the presence DI of CP regions in the CP in the absence of additional DA was obtained, according to which, $B^{(P)}$ and $B^{(K)}$ – $q \times q$ – blocks of the CP matrix F DI for which $q \geq 16$ and $B^{(P)} \cap B^{(K)} = \emptyset$, then in order for $B^{(P)}$, $B^{(K)}$ to be corresponding blocks from the CP display in digital images, it is necessary and sufficient to fulfil (2).

Taking into account the obtained criterion, in order to reduce the possibility of false alarms when using the developed method for detecting CP in the control system, the $q \times q$ – blocks for which are further used $q \geq 16$. Taking into account the results of the experiment and the criterion obtained, at this stage of the study, it is assumed that if post-processing of the DI after cloning took place, then the areas of the CP that were identical before it will change less relative to each other than those that differed from the CP and therefore differed from each other before post-processing.

The work quantitatively determines the characteristic of 'block differences' taking into account (3), i.e., taking into account the introduced characteristic of the volume of the corresponding body. Let B_{ij}, $i = \overline{1, n-q+1}$, $j = \overline{1, m-q+1}$, – $q \times q$ –block of the $n \times m$ – matrix of the central system F, for f_{ij} which the element is located at position (1,1). Let us call this type of block B_{ij} the main. Each block B_{ij} is assigned $(n - q + 1) \times (m - q + 1)$ matrix of block differences $M^{(ij)}$ (MBD) with elements $m^{(i,j)}{}_{k,l}$ defined as:

$$m^{(i,j)}{}_{k,l} = \sum_{t,p=1}^{q} r_{tp},\ k = \overline{1, n-q+1}, l = \overline{1, m-q+1}, \quad (4)$$

where r_{tp}, $t, p = 1, q$ – elements of the $q \times q$ –matrix R, which is obtained as follows:

$$R = \left| B_{ij} - B_{kl} \right|, \quad (5)$$

where equality (5) is understood in an element-by-element sense. The elements of MBD $M^{(ij)}$ reflect the difference between block B_{ij} and any other block of the matrix DI in the sense of (4)-(5), and therefore determine those blocks that differ least from B_{ij} (in terms of the volumes of the corresponding cylindrical bodies), without intersecting with it, and are sought, given the task under consideration.

The established differences $M^{(ij)}$ in properties in B_{ij} cases where belongs/does not belong to the CP region. $\min M^{(ij)} = m^{(i,j)} = 0$. Small in comparison with other elements $M^{(ij)}$ will also be $m^{(i,j)}{}_{k}$, $k \in \{i-1, i, i+1\}$, $l \in \{j-1, j, j+1\}$. However, the smallness of these elements is not an indicator of the similarity of the blocks corresponding to them in the sense considered in the work. Therefore, for the convenience of further research, the element $m^{(i,j)}{}_{ij}$ of the matrix $M^{(ij)}$ and its 8 closest neighbours are made equal to a number that is large compared to other elements $M^{(ij)}$. The new matrix is denoted $\overline{M}^{(ij)}$ and is called the modified matrix of block differences (MMBD). The minimum element $\overline{M}^{(ij)}$ corresponds to the block B_{kl} of the clone (prototype) if is the B_{ij} – block of the prototype (clone). In this case:

$$\min \overline{M}^{(ij)} = \min \overline{M}^{(kl)}. \quad (6)$$

If is B_{ij} not a CP block, then $min\,\overline{M}^{(ij)}$ in general is not related to $min\,\overline{M}^{(kl)}$ for any arbitrary block B_{kl} that does not intersect B_{ij}.

Taking into account (6), the matrix F DI is matched $(n - q + 1) \times (m - q + 1)$ with the matrix G – the matrix of minimum block differences (MBD), whose element g_{ij}, $i = \overline{1, n-q+1}$, $j = \overline{1, m-q+1}$, reflects the smallest difference in the sense of (4)-(5) of $q \times q$ –block B_{ij} from any other block of the matrix F:

$$g_{ij} = min\,\overline{M}^{(ij)}, i = \overline{1, n-q+1}, j = \overline{1, m-q+1}. \quad (7)$$

The corresponding blocks B_{ij} and B_{kl} CP in G will correspond to coinciding values of global (or local in conditions of significant DA, as shown below) minima:

$$g_{ij} = g_{kl}, \quad (8)$$

where g_{tp} is a global minimum if G, and $g_{tp} = \min\limits_{1 \le i \le n-q+1, 1 \le j \le m-q+1} g_{ij}$ a local minimum if G in there exists such $U(g_{tp})$ for g_{tp} a neighbourhood for that: $\forall g_{ij}, \in (g_{tp})$, $g_{ij} \neq g_{tp}$: $g_{ij} > g_{tp}$. The $U(g_{tp})$ neighbourhood r' of radius is created by elements $g_{t+k, p+l}$, $k, l \in (-r', \ldots, 0, \ldots, r')$.

Thus, the formal parameter found is a convergence index, quantitatively identical for the corresponding blocks of the CP of the DI, including under conditions of additional DA: the value of the global (local) minimum for the MMBD corresponding to the analysed DI. As a quantitative metric for assessing the reliability of perception of the obtained digital images, similar to the approaches widely used in modern scientific publications on digital image processing, computer vision, and information security, the peak signal-to-noise ratio (PSNR) was used in this study. This metric was chosen because of its versatility, ease of calculation, and ability to quantitatively reflect the level of image distortion resulting from the application of disturbing influences (Ali *et al.*, 2022; Hosny *et al.*, 2023). The PSNR metric is based on a comparison of the original (reference) digital image and the processed image and allows the root mean square error between corresponding pixels to be estimated (Khalil *et al.*, 2023). High PSNR values indicate insignificant distortions and preservation of visual quality, whereas a decrease in this indicator corresponds to an increase in noise intensity or information loss. This property makes PSNR a convenient tool for analysing the impact of various disturbing factors, including additive noise, filtering operations, compression artefacts, and other types of signal degradation, on the reliability of visual perception. Moreover, PSNR is a standardised and widely recognised metric that is frequently applied for benchmarking image processing algorithms within a unified evaluation framework. Its use ensures experimental reproducibility and facilitates direct comparison of the obtained results with data reported in related studies. In the context of clone detection in digital images, PSNR is particularly relevant, as it enables the determination of acceptable disturbance thresholds under which visual distortions remain imperceptible to the human observer but may significantly affect the performance of automated detection algorithms. Therefore, the application of PSNR provides a balanced compromise between computational efficiency, interpretability, and practical relevance in the domain of digital forensics and information security.

The PSNR value was calculated according to the generally accepted analytical expression and allowed to quantitatively assess the level of image distortion after the application of disturbing influences and to determine the acceptable limits of parameters under which the visual authenticity of the DI is preserved:

$$PSNR = 10 \cdot lg\left(255^2 \Big/ \left(\frac{1}{mn} \sum_{i,j} \left(f_{ij} - \overline{f}_{ij} \right)^2 \right) \right), \quad (9)$$

where f_{ij}, $i = 1, n, j = 1, m$ – elements of the matrix of the modified digital image (it is assumed that the reliability of the digital image is preserved if PSNR > 37 dB).

During the experimental procedure, a set of reference digital images was subjected to controlled disturbing influences with varying intensity levels. For each disturbance scenario, the PSNR value was calculated in order to quantitatively evaluate the degree of image degradation. The obtained PSNR values were subsequently analysed to identify threshold intervals corresponding to visually imperceptible distortions and critical degradation levels that significantly affect both human perception and automated clone detection performance. To ensure statistical reliability, each experiment was repeated multiple times with different disturbance parameters and initial image samples. The resulting measurements were averaged to minimise the influence of random fluctuations and noise. Additionally, graphical visualisation of PSNR trends was performed to facilitate comparative analysis and interpretation of the experimental results.

Thus, the proposed methodological framework combines quantitative image quality assessment with systematic parameter variation, enabling a comprehensive evaluation of the impact of disturbing influences on digital image perception and clone detection robustness. The proposed methodology for clone detection is based on an iterative interval bisection strategy combined with binary image analysis and geometric object comparison. The main objective of the method is to localise characteristic extrema in the interpolated matrix G and to identify repeated structural patterns that indicate the presence of cloning in a digital image. At the initial stage, the full analysis interval $[a, b] = [T_{min}, T_{max}]$ was divided into two equal sub-intervals. For each sub-interval, a corresponding cross-section of the interpolated surface is generated. These cross-sections are represented as binary digital images, where pixel values falling within the specified intensity range are assigned logical values, allowing spatial structures to be isolated.

Each binary cross-section is then analysed using standard connected-component and contour detection techniques. The algorithm searches for objects that are identical in shape and size and bounded by closed contours. The presence of such objects is used as a criterion for updating the counters k_1 and k_2, which indicate whether characteristic patterns are detected in the lower or upper sub-interval, respectively. If both cross-sections contain identical objects ($k_1 = k_2 = 1$), the algorithm selects the interval associated with smaller geometric object dimensions, as this region typically corresponds to a more precise localisation of cloning artefacts. The selected interval is then further divided in half, and the procedure is repeated. This iterative refinement continues until the detected objects become point-like, meaning that their area is equal to one pixel. At this stage, the spatial localisation accuracy reaches its maximum resolution, and the iteration process is terminated. The occurrence of point objects in both analysed sub-intervals is interpreted as a reliable indicator of cloning presence in the digital image.

Finally, the coordinates of the detected point objects are extracted, enabling precise localisation of both the cloned region and its original prototype. This spatial correspondence provides direct geometric evidence of duplication and allows verification of the detected falsification. The proposed approach ensured robust detection performance under additional disturbing influences, such as noise and compression artefacts, by progressively narrowing the analysis interval and isolating stable structural features. As demonstrated in the experimental results, the method provides high localisation accuracy while maintaining computational efficiency.

RESULTS AND DISCUSSION

When modelling attacks on DI at this stage of research, various types of disturbing influences are used, in particular Gaussian, multiplicative and Poisson Noise, salt-and-pepper noise, lossy compression, as well as filtering using various spatial and frequency filters. These types of distortions are typical for simulating image post-processing after cloning operations and correspond to approaches that are widely used in modern scientific research in this and related fields (Hamza, 2022; Jain, 2024). It is important to note that the "author" of a falsified image is usually interested in ensuring that the result of the forgery does not attract additional attention from the expert during visual analysis. In this regard, post-processing of a cloned image is carried out taking into account the requirements for preserving its visual authenticity. Given the widespread use of lossy compression formats for storing and transmitting digital images, the modified image is highly likely to be saved in this format. The most common format in practical application is Joint Photographic Experts Group (JPEG), which justifies its use as the basis for modelling attacks. The results of additional perturbations should not normally lead to noticeable visual artefacts in the final DI. This necessitates careful selection of noise, filtering and compression parameters that ensure the reliability of image perception after their application. At the same time, in practice, there may be situations where post-processing of a DI leads to a partial deterioration in its visual quality.

Note that the reliability of visual perception of a DI under such disturbing influence is considered to be preserved, although the subjective assessment of image quality by individual observers may differ. When modelling attacks on digital images at this stage of research, the following types of distortions are used as the most common and practically significant: additive Gaussian Noise, Multiplicative Noise, "Salt And Pepper", lossy compression, and filtering using various types of filters (9). To evaluate the influence of different noise models on the reliability of digital image perception, a series of controlled experiments was performed using additive and impulse noise disturbances with varying intensity parameters. The average PSNR values obtained for the experimental image set under Gaussian, multiplicative, Poisson, and salt-and-pepper noise conditions are presented in Table 1. The values that satisfy the visual authenticity criterion (PSNR > 37 dB) are highlighted in bold.

Table 1. Average PSNR values obtained as a result of applying noise disturbances to digital images from the experimental array

Gaussian Noise with zero mathematical expectation				**Multiplicative Noise**				**"Salt And Pepper"**	
D = 0.0001	D = 0.0005	D = 0.001	D = 0.005	D = 0.0005	D = 0.001	D = 0.005	D = 0.01	D = 0.02	D = 0.001
43.8	**39.1**	34.5	28.3	**42.8**	**40.4**	33.2	30.9	26.2	**39.0**

Source: created by the authors

The results presented in Table 1 demonstrate that low-intensity noise disturbances have a limited impact on perceived image quality, whereas increasing noise density leads to a rapid degradation of visual fidelity. In particular, Gaussian noise with variance values up to D = 0.0005 preserves acceptable perceptual quality, while higher variance levels result in PSNR values below the reliability threshold. A similar trend is observed for multiplicative and impulse noise, where increasing disturbance intensity significantly

reduces PSNR. These findings indicate that the robustness of visual perception is highly sensitive to noise magnitude and that only a limited range of disturbance parameters can be considered acceptable for preserving digital image authenticity. In order to analyse the effect of lossy compression on the reliability of digital image perception, additional experiments were conducted using the JPEG format with different quality factor (QF) values. The corresponding average PSNR values obtained for both Tagged Image File Format (TIFF) and JPEG output formats are summarised in Table 2. The results provide quantitative insight into the relationship between compression level and visual distortion.

Table 2. Average PSNR values as a result of resaving a digital image in a lossy format (Jpeg) with different values of the QF

Output format DI	QF							
	95	85	75	65	55	45	35	25
Tif	**49.0**	**46.8**	**45.4**	**44.0**	**43.4**	**42.6**	**41.5**	**39.0**
Jpeg	**50**	**41.6**	**40.3**	36.4	35.5	35.1	34.4	34.1

Source: created by the authors

As shown in Table 2, decreasing the quality factor leads to a gradual reduction in PSNR values, indicating an increase in compression-induced distortions. For high QF values (above 75), the PSNR remains well above the perceptual reliability threshold, ensuring visually acceptable image quality. However, when the quality factor drops below 55, a noticeable degradation occurs, and PSNR approaches the critical boundary of visual authenticity. These results confirm that aggressive lossy compression significantly affects the reliability of digital image perception and may compromise the effectiveness of subsequent image analysis and forensic procedures. To assess the impact of spatial filtering operations on digital image quality, experiments were performed using averaging, Gaussian, and median filters with different kernel sizes. The corresponding average PSNR values obtained after filtering are presented in Table 3. This analysis allows the evaluation of smoothing effects on image fidelity and perceptual reliability.

Table 3. Average PSNR values as a result of filtering a digital image

Averaging filter of size $p \times p$			Gaussian filter of size $p \times p$ ($sig = 0.5$)			Median filter 3×3
$p = 3$	$p = 5$	$p = 7$	$p = 3$	$p = 5$	$p = 7$	
37	32	23	**44**	**44**	**44**	**38**

Source: created by the authors

The data presented in Table 3 indicate that filter type and kernel size play a critical role in determining the level of image distortion. Averaging filters with increasing window size lead to a substantial decrease in PSNR due to excessive smoothing and loss of fine structural details. In contrast, Gaussian filtering with moderate kernel sizes demonstrates higher PSNR stability, preserving visual quality more effectively. The median filter exhibits intermediate performance, reducing impulse noise while maintaining acceptable perceptual fidelity. These results confirm that inappropriate filtering parameters may significantly degrade visual authenticity, whereas carefully selected configurations can preserve reliable image perception. In further modelling of disturbing influences on digital images, "Salt And Pepper" noise will not be used. This is due to the fact that even with acceptable PSNR values (Table 1), the superimposition of such noise leads to the appearance of clearly visible visual artefacts on the disturbed image. In addition, the specificity of this type of noise – the presence of separately located point distortions – does not provide reliable masking of cloning results at low values of the noise intensity parameter d. Accordingly, from a practical point of view, this type of disturbance is highly unlikely to be used by the "Author" of DI falsification. A similar conclusion regarding the presence of noticeable visual artefacts can be made for Poisson noise, which is confirmed by the results of visual analysis showing superimposition without localisation of point objects (Fig. 1).

Based on this, at this stage of the study, the following types of transformations were selected as disturbing influences on digital images: Gaussian Noise with zero mathematical expectation and dispersion $D \le 0.0005$ ($D \in \{0.0005; 0.0001\}$); Multiplicative Noise with variance $D \le 0.001$ ($D \in \{0.0005; 0.001\}$); lossy compression with an arbitrary quality factor QF in the range from 25 to 95 filtering, except for an averaging filter with a mask size $p \ge 5$. Regarding the disruptive effect associated with lossy compression of digital images, it should be noted that although for quality factor values $QF < 65$, the PSNR index for images in JPEG format decreases and becomes less than 37 dB, no visually noticeable artefacts are observed. Given that compression is the most common type of post-processing of DI and does not usually attract the attention of experts, compression options with $QF < 65$ are not excluded from further consideration at this stage of the study, despite the formal decrease in quantitative quality indicators. To study the complex of disturbing influences on DI after cloning,

noise with the parameters defined above was superimposed, followed by lossy compression of the images. The compression quality coefficients QF ∈ {65, 75, 85, 95} were selected, since these parameters are the most common and probable from the point of view of practical use during the storage and transmission of DI.

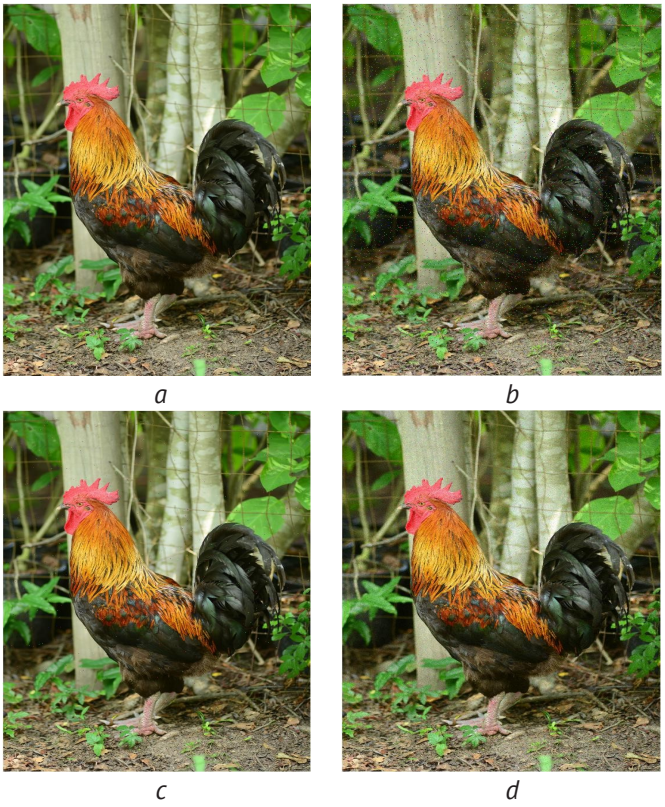


Figure 1. Results of noise superimposition on the DI
Notes: *a* – original DI; *b* – DI after superimposition of "Salt And Pepper" noise with *d* = 0.02; c – "Salt And Pepper" noise with *d* = 0.001; *d* – Poisson Noise
Source: created by the authors

To investigate the combined impact of multiple disturbing factors on digital image quality, an additional series of experiments was performed in which noise disturbances were applied first, followed by lossy JPEG compression with different quality factor values. The resulting PSNR values obtained under complex disturbance conditions are presented in Table 4. This analysis allows the evaluation of cumulative degradation effects and their influence on the reliability of visual perception.

Table 4. PSNR values (dB) after applying complex disturbing effects to a digital image

Output format DI	Gaussian Noise								Multiplicative Noise							
	D = 0.0001				D = 0.0005				D = 0.0005				D = 0.001			
	Subsequent saving in Jpeg with a quality factor QF								Subsequent saving in Jpeg with a quality factor QF							
	QF = 95	QF = 85	QF = 75	QF = 65	QF = 95	QF = 85	QF = 75	QF = 65	QF = 95	QF = 85	QF = 75	QF = 65	QF = 95	QF = 85	QF = 75	QF = 65
Tif	**43.6**	**42.3**	**41.9**	**41.1**	**39.0**	**38.8**	36.9	36.5	**42.7**	**42.5**	**42.3**	**40.0**	**40.3**	**40.0**	**39.5**	**38.7**
Jpeg	**43.7**	**38.9**	**37.2**	36.3	36.9	36.7	36.6	36.2	**41.9**	**38.5**	**37.0**	36.1	**38.7**	**37.5**	36.7	36.0

Source: created by the authors

The results presented in Table 4 demonstrate that the combined application of noise and lossy compression produces a significantly stronger degradation effect compared to the influence of individual disturbances. Even for relatively low noise variance values, subsequent JPEG compression leads to a noticeable reduction in PSNR, indicating cumulative distortion accumulation. In the case of Gaussian noise, high quality factor

values (QF = 95 and QF = 85) still maintain PSNR values above the perceptual reliability threshold, whereas further reduction of the quality factor results in values approaching or falling below the critical boundary. A similar trend is observed for multiplicative noise, where increasing noise intensity combined with compression substantially reduces visual fidelity. Furthermore, the comparison between TIFF and JPEG output formats shows that uncompressed or lossless storage formats demonstrate higher robustness against cumulative distortions, while JPEG images exhibit greater sensitivity to compound disturbing effects. This confirms that the sequence of image processing operations and the choice of output format play a crucial role in preserving visual authenticity.

Overall, the obtained results indicate that complex disturbing effects significantly narrow the acceptable parameter range required to maintain reliable digital image perception. These findings are particularly important for digital forensics applications, where multiple processing stages are often applied sequentially and may collectively compromise both visual integrity and the performance of automated detection algorithms. Taking into account the results of subjective quality ranking obtained in the process of applying a complex of disturbing influences to DI, as well as considering the practically reliable fact of storing DIs in lossyy compression formats, even in cases where the PSNR value is less than 37 dB, within the scope of this study, it is considered appropriate to consider the analysed sets of additional disturbing influences (Table 4) as realistically possible scenarios when creating falsified images by cloning.

For the further development of a method for detecting cloning results under the influence of additional disturbing factors, it is important to have information on the quantitative indicators of the difference between the corresponding blocks of the clone and its prototype after post-processing of the DI. In order to determine these indicators, a computational experiment was carried out using digital images from the experimental array. During the experiment, digital images were cloned, with the cloned area and the corresponding original area being square blocks of size $q \times q$, where $q \in \{32, 24, 16\}$. After cloning, Gaussian or Multiplicative Noise was applied to the images with parameters determined in the previous stages of the study as ensuring the reliability of the visual perception of the DI. Next, the images were saved in lossyless (TIF) and lossy (JPEG) formats with different quality coefficients $QF \in \{65, 75, 85\}$. For the images obtained in this way, taking into account that the clone block was denoted as $B^k = B_{ij}$, and the corresponding prototype block as $B^P = B_{kl}$, matrices $\bar{M}^{(ij)}$ and $\bar{M}^{(kl)}$ were constructed. For each pair of such matrices, the minimum values of the selected difference measure were determined, as well as the coordinates of their location. This made it possible to quantitatively assess the degree of similarity between the clone and prototype blocks after applying disturbing influences and post-processing. The results of the computational experiment are presented in Table 5, in which the following designations are used: I – the average value of the indicator for all DI of the experimental array $\min^{(ij)} = \min^{(kl)}$; II and III – the minimum and maximum values of the indicator among all DI involved in the experiment, respectively $\min^{(ij)} = \min^{(kl)}$.

Table 5. Quantitative estimates of the difference between the clone and prototype blocks in terms of (1) – (2) under image post-processing conditions

Block size	Noise type	Noise parameters	Output format DI after cloning		I	II	III
32 × 32	Gaussian noise	D = 0.0001	Tif		2930	2208	3104
			Jpeg	QF = 85	4068	2183	5764
				QF = 75	4357	2113	6764
				QF = 65	4680	1900	7375
		D = 0.0005	Tif		6544	4561	7138
			Jpeg	QF = 85	7242	4429	8592
				QF = 75	6656	4177	9458
				QF = 65	6799	3900	10147
	Multiplicative noise	D = 0.0005	Tif		2650	408	6258
			Jpeg	QF = 85	3930	1169	6231
				QF = 75	4240	1142	7661
				QF = 65	4572	1252	7603
		D = 0.001	Tif		3742	707	7770
			Jpeg	QF = 85	4796	1240	7536
				QF = 75	4941	1370	8572
				QF = 65	5207	1446	8807

Continued Table 5.

Block size	Noise type	Noise parameters	Output format DI after cloning		I	II	III
24 × 24	Gaussian noise	D = 0.0001	Tif		1649	1277	1753
			Jpeg	QF = 85	2325	1181	3670
				QF = 75	2549	1094	4307
				QF = 65	2746	1068	4369
		D = 0.0005	Tif		3687	2593	3994
			Jpeg	QF = 85	4071	2661	4970
				QF = 75	3946	2400	5467
				QF = 65	3889	2080	5940
	Multiplicative noise	D = 0.0005	Tif		1472	309	4917
			Jpeg	QF = 85	2245	701	3470
				QF = 75	2368	634	4261
				QF = 65	2597	711	4290
		D = 0.001	Tif		1967	501	4209
			Jpeg	QF = 85	2564	711	4095
				QF = 75	2671	731	4511
				QF = 65	2690	801	4712
16 × 16	Gaussian noise	D = 0.0001	Tif		911	501	1123
			Jpeg	QF = 85	1007	546	1241
				QF = 75	1111	678	1535
				QF = 65	1247	654	1987
		D = 0.0005	Tif		1698	901	2145
			Jpeg	QF = 85	1814	959	2297
				QF = 75	1746	811	2611
				QF = 65	1800	875	2849
	Multiplicative noise	D = 0.0005	Tif		662	60	1360
			Jpeg	QF = 85	1068	322	1717
				QF = 75	1229	344	1981
				QF = 65	1338	355	2467
		D = 0.001	Tif		1043	235	2290
			Jpeg	QF = 85	1376	401	2211
				QF = 75	1398	484	2268
				QF = 65	1478	489	2329

Source: created by the authors

Based on the data presented in Table 5, the threshold values $T_{min}^{(q)}$ and $T_{max}^{(q)}$ were determined, corresponding, respectively, to the minimum and maximum levels of difference between a clone block of size $q \times q$ and the corresponding prototype block. When solving the problem of finding areas (blocks) of a digital image with minimal differences between each other, each block B_{ij} is assigned a block difference matrix of size $(n-q+1) \times (m-q+1)$, which is further referred to as MAMBD. The elements of this matrix m^{ij}_{kl} are defined as a measure of the difference between block B_{ij} and all other blocks B_{kl} of the analysed digital image:

$$m^{ij}{}_{kl} = \sum_{t,p}^{g} r_{tp},\ k = \overline{1, n - q + 1},\ l = \overline{1, m - q + 1}, \quad (10)$$

where r_{tp}, $t, p = \overline{1, q}$, are elements of the $q \times q$ matrix R, obtained as follows:

$$R = |B_{ij} - B_{kl}|. \quad (11)$$

In this case, relation (11) is interpreted in an element-by-element sense, i.e., comparison operations are performed for the corresponding pixels of the blocks, which allows for a correct assessment of their similarity after cloning and possible disturbing influences. The obtained threshold values were determined for three levels of the parameter q. In particular, when q equals 32, the minimum threshold value $T_{min}(32)$ is 408, while the corresponding maximum threshold value $T_{max}(32)$ reaches 10,147. For the case where q equals 24, the minimum threshold $T_{min}(24)$ is 309 and the maximum threshold $T_{max}(24)$ is 5,940. Finally, when q equals 16, the minimum threshold $T_{min}(16)$ is equal to 60, whereas the maximum threshold $T_{max}(16)$ amounts to 2,849. These threshold values quantitatively characterise the range of possible differences between the corresponding blocks of the clone and the prototype after cloning and subsequent post-processing of the DI. In addition, it is possible to determine similar threshold values

separately for each type of disturbing influence, which is consistent with the approaches presented in S. Grygorenko (2016). In order to formalise the process of detecting cloning results in DI and ensure the possibility of their quantitative analysis, it is advisable to move from the direct analysis of pixel values to the study of specially constructed mathematical objects. This approach allows us to generalise information about local similarities between image fragments, reduce the influence of random noise distortions, and increase the method's resistance to additional disturbing influences. In this context, a key step is to construct an auxiliary matrix that accumulates information about the degree of difference between all possible blocks of a fixed-size DI. Analysis of the properties of this matrix and the nature of its extreme creates a theoretical basis for identifying clone regions and the corresponding prototype, as well as for localising cloning results in the image space.

The main object of analysis during the detection of clone and prototype regions in a DI is the introduction of an $n \times m$ matrix G of size $(n-q+1) \times (m-q+1)$ with elements g_{ij}, where $i = \overline{1, n-q+1}$, $j = \overline{1, m-q+1}$. The values of the elements g_{i_j} are determined according to the relation: $g_{ij} = \min^{(ij)}$, $i = \overline{1, n-q+1}$, $j = \overline{1, m-q+1}$. For this matrix, it is justified that the blocks B_{i_j} and B_{kl}, which correspond to the clone and its prototype, will correspond to the same global (or local) minima in the matrix G, $g_{i_j} = g_{kl}$, which for most DI is an indicator of the presence of cloning and allows its spatial location to be determined. At the same time, in practice, the ratio $g_{i_j} = g_{kl}$ may indicate not the clone and prototype regions, but the original, but slightly different image fragments. In such cases, there is a risk of false cloning identification.

In order to refine the localisation of the clone and prototype areas, it is proposed to use the construction of a set of "sections" of the surface, which is a graph of a function that interpolates the elements of the matrix G. Such "sections" are defined as inequalities $c_1 \le z \le c_2$, where c_1, $c_2 = \text{const}$, $c_1, c_2 \in N$, and N is the set of natural numbers. As a result of performing the specified "sections", the coordinates of those blocks of the DI are determined for which the minimum difference from all other blocks in terms of criteria (10) – (11) falls within the specified interval $[c_1, c_2]$. This approach allows for more accurate localisation of potential cloning areas and reduces the number of false positives in the algorithm.

The "sections" is performed according to a specially developed method, which will hereinafter be referred to as the method of dividing the segment in half for the analysis of the matrix of minimum block differences (MAMBD). The proposed method is based on the idea of iterative narrowing of the range of values of the elements of the minimum block difference matrix in order to localise areas that potentially correspond to cloning results. This approach allows for the effective separation of the clone area and its prototype from other structurally similar but original fragments of the DI (Li *et al.*, 2022; Jaiswal *et al.*, 2022). The MAMBD method is an integral part of the basic method of detecting cloning results in DI presented below and performs the function of a refinement stage of analysis. Its application is aimed at increasing the stability of detection against additional disturbing influences, such as noise, filtering and lossy compression, as well as reducing the number of false positives. By using a split-half interval procedure, the method provides adaptive selection of threshold values corresponding to the most informative areas of the surface that interpolates the elements of the minimum block difference matrix (Koul *et al.*, 2022). The main stages of the MAMBD method, outlined below, form a sequential analysis procedure that covers the selection of block representation parameters, the determination of difference thresholds, and the construction of corresponding "sections" for localising potential cloning areas. Together, this creates a methodological basis for the further algorithmic implementation of the method for detecting cloning of DI in post-processing conditions and under the influence of disturbing factors.

Taking into account the specifics of the analysed DI of size $n \times m$ or the geometric parameters of the area "suspicious" from the point of view of cloning, select the size of the analysed block $q \times q$. Then, for the DI, construct a matrix G of size $(n-q+1) \times (m-q+1)$, which corresponds to the MAMBD. Determine the threshold values Tmin and Tmax for the possible values of the elements of the matrix G that satisfy the relation $g_{ij} = g_{kl}$, i.e., correspond to the blocks of the clone and its prototype. Hereafter, the element gtp is considered to be the global minimum of the matrix G if it corresponds to the smallest value among all elements within the analysed region. The element gtp is defined as a local minimum if there exists a neighbourhood $U(g_{tp})$ in the matrix G such that, for any element $g_{ij} \in U(g_{tp})$, where $g_{ij} \ne g_{tp}$, the inequality $g_{ij} > g_{tp}$ holds. The neighbourhood $U(g_{tp})$ with radius r is formed by the elements $gt+k, p+l$, where k, $l \in \{-r, \dots, 0, \dots, r\}$ k, which enables the analysis of the local region surrounding the considered element.

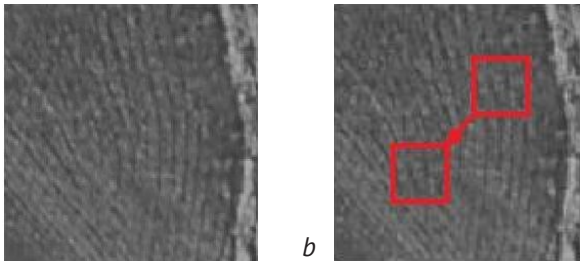
Given a fixed and uniquely specified block size $q \times q$, the threshold values are taken as $T_{min} = T_{min}^{(q)}$ and $T_{max} = T_{max}^{(q)}$. In the case when additional disturbing influences applied to the DI after cloning are known, the threshold values T_{min} and T_{max} are determined based on the experimental data given in Table 5. If information about the nature of the disturbing influences is missing and the block size q cannot be determined unambiguously, it is advisable to use universal estimates as threshold values to ensure the robustness of the method to uncertainty in the analysis conditions: $T_{min} = \min\limits_{g} T_{min}^{(q)}$, $T_{max} = \min\limits_{g} T_{max}^{(q)}$. Procedure for dividing the segment in half. At this stage, an iterative division of the analysis interval is performed in order to refine the parameter range and localise characteristic objects in the digital image. First, the initial bounds of the analysis interval are defined: the lower bound $a = T_{min}$ and the upper

bound $b = T_{max}$. Then, the midpoint of the interval is calculated as $c = (a + b) / 2$. After determining the midpoint, two cross-sections of the surface representing the graph of the function interpolating the elements of matrix G are generated. The first cross-section corresponds to the interval $a \le z \le ca$, while the second corresponds to $c \le z \le b$. The obtained cross-sections are represented as binary digital images, where white pixels indicate those elements of matrix *G* whose values fall within the corresponding intervals.

Next, the counters $k_1 = 0$ and $k_2 = 0$ are initialised and used to record the presence of characteristic objects in the respective cross-sections. The generated binary images are then analysed using standard digital image processing techniques to detect objects bounded by closed contours. If objects of identical shape and size are detected in the cross-section $a \le z \le c$, the counter k_1 is incremented by one. If, in addition, the area of such objects equals one, meaning that the objects are point-like and isolated, the algorithm proceeds to the next step.

A similar procedure is applied to the cross-section $c \le z \le b$. If objects with identical shape and size are detected, the counter k_2 is incremented. If the area of these objects is equal to one, the algorithm also proceeds to the next step. In the case where $k_1 = 1$ and $k_2 = 0$, the cross-section $a \le z \le c$ is selected for further analysis. The upper bound of the interval is updated as $b = c$, after which a new midpoint is calculated and the procedure returns to the next iteration. If the condition $k_1 = 0$ and $k_2 = 1$ is satisfied, the cross-section $c \le z \le b$ is selected for further analysis. In this case, the lower bound is updated as $a = c$, the new midpoint is calculated, and the next iteration is performed. When $k_1 = k_2 = 1$, the cross-section containing objects with smaller geometric dimensions is selected. If the interval $a \le z \le c$ is chosen, the upper bound is set to $b = c$; otherwise, the lower bound is set to $a = c$. Subsequently, a new midpoint is calculated and the procedure continues. If $k_1 = k_2 = 0$ and, at the same time, the condition $c - a > P$ is satisfied, where P is a threshold value determined experimentally, both cross-sections are subjected to further analysis. For each of them, new interval bounds are defined separately, after which the bisection procedure is repeated. Thus, the interval bisection procedure ensures a gradual narrowing of the search region and increases the accuracy of localising characteristic structures in the digital image, which is a necessary condition for the effective implementation of subsequent stages of the algorithm.

If the condition $k_1 = k_2 = 0$ is true for each of the constructed cuts, then it is concluded that the analysed DI has not been cloned. If $k_1 = 1$ or $k_2 = 1$, it is considered that the analysed DI has been cloned. In this case, the corresponding blocks of the clone and its prototype, which are characterised by the minimum difference between each other in terms of criteria (10) – (11), are determined by isolated identical white objects with an area of 1 (one pixel) on the last binary image slice in order. As methods for processing DI, standard approaches to contour extraction can be applied, in particular the methods proposed and implemented in R. Kundu (2022), Lumenci (2024). For example, gradient methods using the color grading function can be used. During experimental studies using DI from the experimental array, it was found that the threshold value of parameter *P* is 300. To illustrate the proposed method, consider a DI that has been cloned and then subjected to Gaussian Noise with zero mathematical expectation and variance D = 0.0001, and saved in JPEG format with a quality factor QF = 75 (Fig. 2, b).



a *b*

Figure 2. DI used to illustrate the operation of MAMBD under conditions of additional disturbing influences
Notes: *a* – original DI; *b* – DI resulting from cloning under conditions of additional disturbing influences
Source: created by the authors

For comparison, the corresponding original DI is shown (Fig. 2). It should be noted that the image under study was deliberately chosen without clearly defined objects, which in general complicates the task of detecting cloning and emphasises the robustness of the proposed approach in conditions of low image structure. First consider a falsified DI (Fig. 2). For its analysis, will use medium-sized blocks with the parameter $q = 24$. According to the data in Table 5, for the selected block size, have the following threshold values: $T_{min} = 1{,}094$, $T_{max} = 4, 307$. Figure 3 shows "cross-sections" in the form of binary images obtained as a result of the first division of the segment $[a, b] = [T_{min}, T_{max}]$ in half. Analysis shows that both sections contain objects of the same shape and size within their boundaries, bounded by closed curves, and therefore $k_1 = k_2 = 1$.

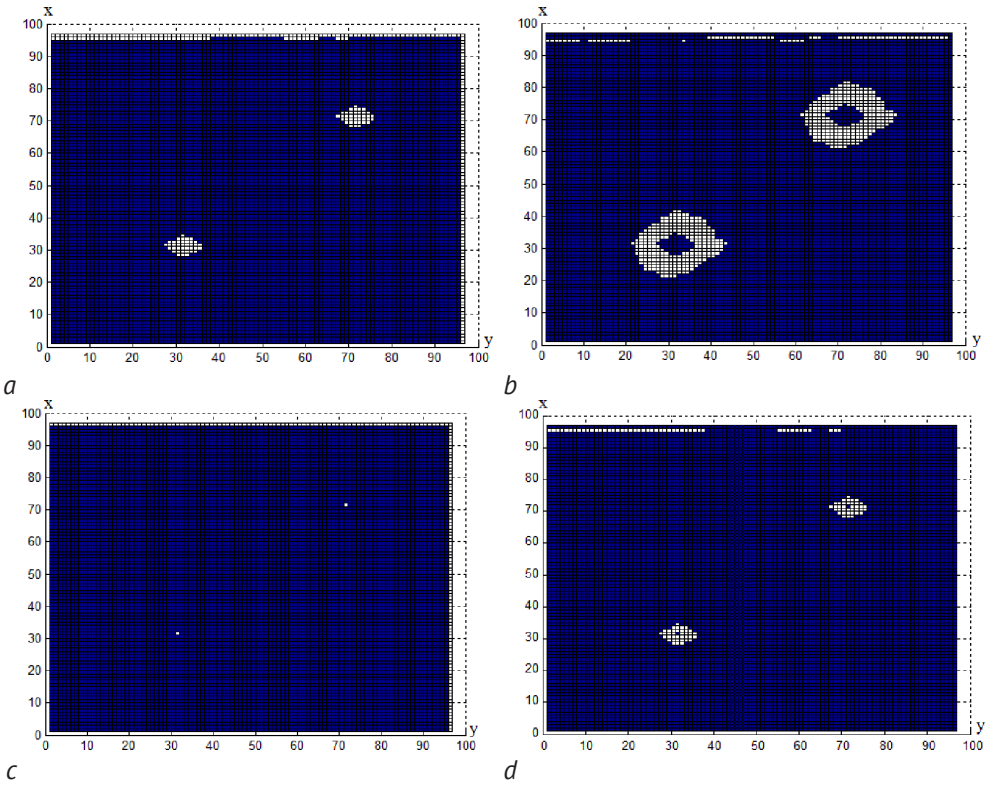


Figure 3. Sequence of sections for MAMBD DI subjected to cloning under conditions of additional disturbances
Notes: a -1,094 ≤ z ≤ 2,700; b – 2,701 ≤ z ≤ 4,307; c – 1,094 ≤ z ≤ 1,897; d – 1,898 ≤ z ≤ 2,700
Source: created by the authors

In accordance with step 3 of the proposed method, the section 1,094 ≤ z ≤ 2,700 is selected for further analysis (Fig. 3). Next, the segment [1,094, 2,700] is again divided in half, after which two more "sections" are constructed (Fig. 3). According to the criteria of the method, the interval 1,094 ≤ z ≤ 1,897 is selected for further consideration. At this stage, the iterative analysis process is completed, since the area of identical objects in the corresponding binary image is equal to 1, i.e., the objects are point objects. As a result, given that $k_1 = k_2 = 1$, it is concluded that cloning is present in the analysed DI. In addition, the spatial location of the clone and its prototype is determined by isolated point objects, namely blocks with coordinates (31, 31) and (71, 71), which fully corresponds to the actual nature of the falsification (Fig. 3). Considering the original DI (Fig. 2), using the same parameter values for analysis as in the case of the falsified image (although in general this condition is not mandatory). Figure 4 shows the "sections" in the form of binary images obtained as a result of the first division of the interval $[a, b] = [T_{min}, T_{max}]$ in half.

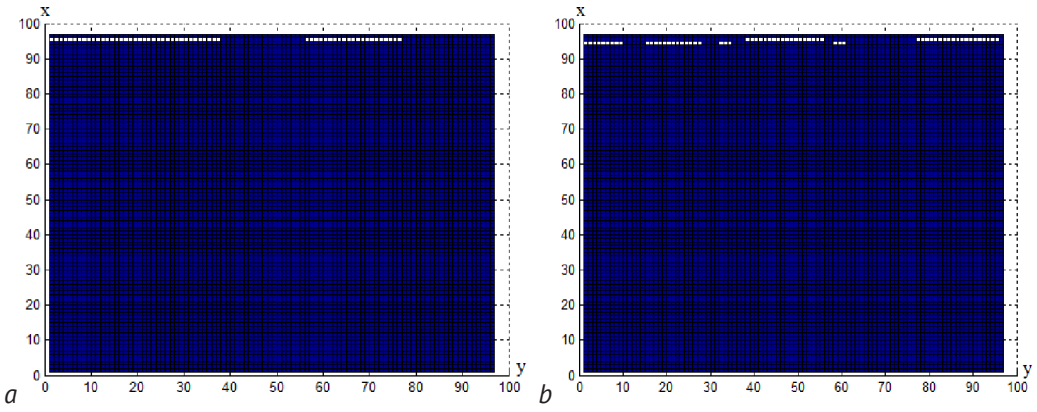


Figure 4. Sequence of cross-sections for the original DI
Notes: a -1,094 ≤ z ≤ 2,700; b – 2701 ≤ z ≤ 4,307
Source: created by the authors

Further iterative refinement of the segment did not lead to qualitative changes in the form of the obtained sections, therefore the corresponding images are not shown. Analysis of the constructed binary images showed that no identical objects bounded by closed curves were found within any of the "sections". Accordingly, at this stage, the value $k_1 = k_2 = 0$ is recorded. At step 4 of the algorithm, based on the results obtained, a conclusion is made about the absence of cloning in the analysed DI, which is fully consistent with its actual status as original. The proposed MAMBD method of analysing the minimum block difference matrix does not in all cases provide an unambiguous clarification of the presence and spatial location of the clone and its prototype. The complexity of obtaining the result is due to the fact that the neighbourhoods of the clone and the prototype on the corresponding "sections" may not form closed curves in general. Moreover, in certain situations, such regions may represent discrete sets of points or short segments, which complicates their automated selection using standard DI processing methods.

As a result, the effectiveness of the MAMBD method is reduced for images with a weakly expressed structure or a high level of textural homogeneity, which requires the use of additional analysis criteria or combining the method with other approaches to detecting cloning results. Possible ways to overcome these limitations include combining the MAMBD method with additional criteria for spatial-structural analysis and using multi-level approaches to interpret the results. In particular, it is advisable to use morphological operations (expansion, narrowing, closing) to aggregate discrete points and segments into more connected areas, which facilitates their further selection and analysis. In addition, the stability of the localisation of the clone and prototype areas can be improved by adaptively changing the size of the analysed block q and using multi-scale analysis, which allows different levels of spatial detail in the image to be taken into account. Another promising direction is the integration of statistical and machine classification methods, in particular the use of clustering to group elements of the minimum block difference matrix that are close in value, or the use of machine learning models to distinguish between true and false positive manifestations of cloning. Additionally, the effectiveness of the method can be increased by combining MAMBD with feature-based approaches (based on key points, texture or frequency characteristics), which allows compensating for the shortcomings of each individual method and ensuring more reliable detection of cloning results in complex conditions of DI post-processing.

The results obtained confirm that the reliability of visual perception of DI and the stability of digital examination procedures significantly depend on the type and intensity of disturbing influences, as well as on their combinations. The paper justifies the use of PSNR as a quantitative criterion for controlling the "visual authenticity" of DI (threshold PSNR > 37 dB) and experimentally determines the ranges of noise, JPEG compression and filtering parameters at which distortions remain inconspicuous to humans, but can complicate or mask traces of forgery. This approach is consistent with the general trend in contemporary research in digital forensics in the work of W. Haiwei *et al.* (2022), which identifies not only the fact of manipulation, but also the robustness of methods to post-processing (noise, compression, filtering), which is a typical "anti-forensic" context. In the works of F.Z. Mehrjardi *et al.* (2023) and L.-X. Jiao *et al.* (2025), a significant part of the focus is on deep learning for detecting forgeries, in particular copy-move (cloning) and related manipulations, systematise deep learning approaches (CNN/transformer architectures, hybrids with "classical" features) and emphasise that the main challenge is degradation from real distribution channels (social networks, re-saving, filtering), which change image statistics and reduce the accuracy of detectors. The authors of the paper support the conclusions of these reviews: post-processing is a critical factor, and without an explicit "disturbance model" or at least quantitative quality control (such as PSNR), it is difficult to guarantee the reliability of expert conclusions.

At the same time, the recommended methodology differs from the dominant deep learning solutions in that it is interpretable and training-independent. The authors use an iterative procedure of dividing the interval $[T_{min}, T_{max}]$ in half and analysing the "sections" of the interpolated surface of the G matrix in the form of binary images. The decision on the choice of a subinterval is made based on the presence in both "sections" of objects of the same shape/size and their geometric characteristics (counters k_1, k_2), which allows localising the clone and prototype to point objects (area = 1) and determine the coordinates of the source blocks of the forgery. In practice, this brings our approach closer to a class of detectors that rely on stable structural/geometric manifestations of duplication characteristic of classical forensics (key points, block features), but implement localisation through controlled narrowing of the *z* parameter range rather than through mass comparison of descriptors. This complements the research of N.A.M. Abir *et al.* (2021), G. Suresh & C.S. Rao (2022), L.-X. Jiao *et al.* (2025), which describe keypoint-based and block-based CMFD methods.

The experimental conclusions of the authors of the work regarding the influence of JPEG compression, noise and filtering are consistent with the works of L. Bertojo *et al.* (2022), M. Aloraini *et al.* (2021), G. Tahaoglu *et al.* (2021), which demonstrate the robustness of CMFD to typical attacks (JPEG, Gaussian noise) under certain conditions, but also emphasise the dependence of the results on the intensity of degradation. In fact,

Tables 1 – 4 in the paper show that the "acceptable" disturbance parameters (according to the PSNR criterion) are sharply narrowed under complex influences (noise + JPEG), which logically coincides with practical scenarios of online content distribution: consecutive re-encodings and platform-specific transformations can accumulate distortions. A similar emphasis on degradation from online environments is presented in the work of W. Haiwei *et al.* (2022), which proposes modelling "social media noise" and training robust detectors. Authors introduced a quantitative framework (PSNR threshold) for selecting disturbance modes under which visual authenticity is preserved but the forensic task is complicated. According to the reaserchers, this approach is useful both for evaluating the robustness of algorithms and for standardising experiments.

N. Taneja *et al.* (2025), C. Shi *et al.* (2023), H. Xie *et al.* (2022) emphasise that contrast enhancement, median filtering, or other transformations can be deliberately applied to mask traces of manipulation, and therefore detection methods must either be resistant to them or have quality control/change procedures in place. Observations by aforementioned authors on the significant impact of filtering (especially averaging with large windows) on PSNR and, accordingly, on "perceptual reliability" and potential detectability of cloning, and that filtering can "smooth out" local structures, which is critical for methods that rely on geometry/texture. J. Yang *et al.* (2021) proposed an algorithm based on key points or two-stage feature matching schemes and confirm that realistic attacks (scaling, rotations, noise, JPEG) are effective. The authors of the work emphasise that the difference between this approach and others is that instead of searching for a large number of matches, use iterative "sections" analysis and the criterion of minimising the geometric dimensions of objects as a convergence marker, which allows us to localise the clone/prototype in block coordinates and work in conditions of additional distortions controlled by PSNR. In summary, this work confirms the general conclusions of the current literature regarding the decisive role of post-processing and degradation in digital forensics tasks, but offers a more interpretable and parametrically controlled approach: the selection of acceptable perturbations through the PSNR threshold and the localisation of cloning through iterative interval division and analysis of binary "sections". At the same time, the practical value lies in the fact that this approach can be used as a standalone detection procedure or as a robustness assessment module/protocol for other CMFD methods (including deep learning) that have been actively developing in recent years.

CONCLUSIONS

As a result of the study, an urgent scientific and applied problem related to the development of a basic method for detecting cloning results in DI under the influence of additional disturbing factors has been solved. The proposed approach was based on the analysis of the minimum block difference matrix and the application of the segment splitting method, which provides a formalised procedure for detecting and localising the areas of the clone and its prototype. The paper justified the choice of types and parameters of disturbing influences that are most likely from a practical point of view and do not violate the reliability of visual perception of digital images. Based on computational experiments, quantitative threshold values for differences between corresponding clone and prototype blocks for different block sizes have been determined, which creates the basis for further algorithmic implementation of the method. It has been shown that the proposed algorithm allows not only to establish the fact of cloning, but also to determine the spatial location of the corresponding areas, while maintaining resistance to noise, filtering and lossy compression.

Further research should be directed towards improving the proposed method by integrating additional spatial-structural analysis criteria and morphological operations that allow discrete manifestations of cloning to be aggregated into connected regions. The development of a multi-scale approach with adaptive selection of the size of the analysed blocks depending on the local properties of the image, as well as the combination of MAMBD with feature-based, statistical and machine learning methods of analysis in order to reduce the number of false positive decisions, is promising. Of particular interest is the expansion of the class of studied disruptive influences, including complex combined attacks and geometric transformations, as well as the practical testing of the method in the tasks of digital forensics and information security. The implementation of these directions will increase the versatility, robustness and applied value of the proposed approach and create the prerequisites for the development of intelligent systems for the automated detection of DI falsification.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Abir, N.A.M., Abd Warif, N. B., & Zainal, N. (2021) An evaluationofpatch match-based copy-move forgery detection (CMFD) onsocialmedia images. In *2021 IEEE international conference on artificial intelligence in engineering and technology (IICAIET)* (pp. 1-6). Kota Kinabalu: IEEE. doi: 10.1109/IICAIET51634.2021.9573571.

[2] Akram, A., Arfan, M., Rashid, J., Mahmood, K., & Ghani, A. (2025) Advanced digital image forensics: A hybrid framework for copy-move forgery detection in multimedia security. *Journal of Forensic Sciences*, 70 (5), 1801-1823. doi: 10.1111/1556-4029.70076.
[3] Ali, S.S., Ganapathi, I.I., Vu, N.-S., Ali, S.D., Saxena, N., & Werghi, N. (2022) Image forgery detection using deep learning by recompressing images. *Electronics*, 11(3), article number 403. doi: 10.3390/electronics11030403.
[4] Aloraini, M., Sharifzadeh, M., & Schonfeld, D. (2021) Sequential and patch analyses for object removal video forgery detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3), 917-930. doi: 10.1109/tcsvt.2020.2993004.
[5] Bertojo, L., Néraud, C., & Puech, W. (2022) A very fast copy-move forgery detection method for 4K ultra HD images. *Frontiers in Signal Processing*, 2, article number 906304. doi: 10.3389/frsip.2022.906304.
[6] Duszejko, P., Walczyna, T., & Piotrowski, Z, (2025) Detection of manipulations in digital images: A review of passive and active methods utilizing deep learning. *Applied Sciences*, 15(2), 881. doi: 10.3390/app15020881.
[7] Gani, G., & Qadir, F. (2020) A robust copy-move forgery detection technique based on discrete cosine transform and cellular automata. *Journal of Information Security and Applications*, 54, article number 102510. doi: 10.1016/j.jisa.2020.102510.
[8] Ghai, A., Kumar, P., & Gupta, S. (2024) A deep-learning-based image forgery detection framework for controlling the spread of misinformation. *Information Technology & People*, 37(2), 966-997. doi: 10.1108/ITP-10-2020-0699.
[9] Grygorenko, S.N. (2016) Development of a method for detecting cloning in digital images under conditions of additional disruptive actions. *Legal, Regulatory and Metrological Support for the Information Protection System in Ukraine*, 1(31), 85-98.
[10] Haiwei, W., Zhou, J., Tian, J., & Liu, J. (2022) Robust image forgery detection over online social network shared images. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 13430-13439). New Orleans: IEEE. doi: 10.1109/CVPR52688.2022.01308.
[11] Hamza, A. (2022) *How to make salt pepper noise own code*. Retrieved from https://surl.li/empafc.
[12] Hosny, K.M., Mortda, A.M., Lashin, N.A., & Fouda, M.M. (2023) A new method to detect splicing image forgery using convolutional neural network. *Applied Sciences*, 13(3), article number 1272. doi: 10.3390/app13031272.
[13] Jain, A. (2024) *Salt and pepper noise and how to remove it in machine learning*. Retrieved from https://surl.lu/nezlap.
[14] Jaiswal, A.K., & Srivastava, R. (2022) Detection of copy-move forgery in digital image using multi-scale, multi-stage deep learning model. *Neural Processing Letters*, 54(1), 75-100. doi: 10.1007/s11063-021-10620-9.
[15] Jiao, L.-X., Ng, K.-W. & Tong, H.-L. (2025) A comprehensive review of keypoint-based copy-move forgery detection: Challenges and advances. *IEEE Access*, 13, 180941-180952. doi: 10.1109/ACCESS.2025.3618716.
[16] Kadam, K.D., Ahirrao, S., & Kotecha, K. (2022) Efficient approach towards detection and identification of copy move and image splicing forgeries using mask R-CNN with mobilenet V1. *Computational Intelligence and Neuroscience*, 1, article number 6845326. doi: 10.1155/2022/6845326.
[17] Khalil, A.H., Ghalwash, A.Z., Elsayed, H.A., Salama, G.I. & Ghalwash, H.A. (2023) Enhancing digital image forgery detection using transfer learning. *IEEE Access*, 11, 91583-91594. doi: 10.1109/ACCESS.2023.3307357.
[18] Koul, S., Kumar, M., Khurana, S.S., Mushtaq, F. ,& Kumar, K. (2022) An efficient approach for copy-move image forgery detection using convolution neural network. *Multimedia Tools and Applications*, 81(8), 11259-11277. doi: 10.1007/s11042-022-11974-5.
[19] Kundu, R. (2022) *Image processing: Techniques, types & applications*. Retrieved from https://www.v7labs.com/blog/image-processing-guide.
[20] Li, F., Pei, Z., Wei, W., Li, J., & Qin, C. (2022) Image forgery detection using tamper-guided dual self-attention network with multiresolution hybrid feature. *Security and Communication Networks*, 1, article number 1090307. doi: 10.1155/2022/1090307.
[21] Lumenci. (2024) *Digital image processing (DIP): A beginner's guide*. Retrieved from https://surli.cc/qwhewc.
[22] Mehrjardi, F.Z., Latif, A.M., Zarchi, M.S., & Sheikhpour, R. (2023) A survey on deep learning-based image forgery detection. *Pattern Recognition*, 144, article number 109778. doi: 10.1016/j.patcog.2023.109778.
[23] Shehin, A.U., & Sankar, D. (2024) Copy move forgery detection and localisation robust to rotation using block based discrete cosine transform and eigenvalues. *Journal of Visual Communication and Image Representation*, 99, article number 104075. doi: 10.1016/j.jvcir.2024.104075.
[24] Shi, C., Chen, L., Wang, C., Zhou, X., & Qin, Z. (2023). Review of image forensic techniques based on deep learning. *Mathematics*, 11(14), article number 3134. doi: 10.3390/math11143134.
[25] Suresh, G., & Rao, C.S. (2022) Detection and localization of copy-moveforgery in digital images: Review and challenges. *International Journal of Image and Graphics*, 23(4), article numer 2350025. doi: 10.1142/S0219467823500250.

[26] Tahaoglu, G., Uluas, G., & Ustubioglu, B. (2021) A new approach for localization of copy-move forgery in digital images. In *44th international conference on telecommunications and signal processing (TSP)* (pp. 183-186). Brno: IEEE. doi: 10.1109/TSP52935.2021.9522680.
[27] Taneja, M.E.N., Mishra, G.S., & Bhardwaj, D. (2025) Unmasking anti-forensic techniques: A DCNN-driven approach to uncover contrast enhancement and median filtering detection. *Journal of Forensic Sciences*, 70(6), 2324-2337. doi: 10.1111/1556-4029.70161.
[28] Xie, H., Ni, J., & Shi, Y.Q. (2022). Dual-domain generative adversarial network for digital image operation anti-forensics. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3), 1701-1706. doi: 10.1109/TCSVT.2021.3068294.
[29] Yang, J., Liang, Z., Gan, Y., & Zhong, J. (2021). A novel copy-move forgery detection algorithm via two-stage filtering. *Digital Signal Processing*, 113, article number 103032. doi: 10.1016/j.dsp.2021.103032.

Метод виявлення результатів клонування цифрових зображень в умовах постобробки

Світлана Григоренко
Кандидат технічних наук
Національний університет «Одеська політехніка»
65044, пр. Шевченка, 1, м. Одеса, Україна
https://orcid.org/0009-0006-4551-8243
Олександр Євсеєв
Старший викладач
Національний університет «Одеська політехніка»
65044, пр. Шевченка, 1, м. Одеса, Україна
https://orcid.org/0009-0009-3100-2643
Микола Воронов
Аспірант
Національний університет «Одеська політехніка»
65044, пр. Шевченка, 1, м. Одеса, Україна
https://orcid.org/0009-0003-2816-0495

Анотація. Актуальність цього дослідження обумовлена зростанням кількості випадків фальсифікації цифрових зображень та поширеним використанням клонування фрагментів як поширеної техніки маніпуляції, що створює серйозні виклики для інформаційної безпеки. Метою цього дослідження було розроблення та експериментальне підтвердження надійного методу виявлення клонування цифрових зображень в умовах постобробки на основі аналізу матриці мінімальної середньої різниці блоків, що дозволяє надійно ідентифікувати та просторово локалізувати клоновані області та їхні оригінальні прототипи. Запропонований метод базувався на ітеративному бісекційному інтервалі в поєднанні з бінарним перехресним аналізом різниць блоків і був реалізований в середовищі Matlab. Експериментальна структура включала моделювання адитивного гаусового, мультиплікативного та імпульсного шуму, стиснення JPEG, просторове фільтрування та комбіновані спотворення. Візуальна точність була кількісно оцінена за допомогою пікового відношення сигнал/шум із пороговим значенням відношення сигналу до шуму > 37 дБ. Обчислювальні експерименти проводилися на центральних областях цифрових зображень з роздільною здатністю 512×512 пікселів. Клоновані області займали менше 0,85 % площі зображення, зазвичай від 0,098 % до 0,39 %, при цьому також враховувалися менші клоновані області, менше 0,098 %. Аналіз проводився на випадково обраному одноколірному каналі з використанням блокової обробки. За відсутності спотворень після обробки глобальний мінімум функції мінімальної середньої різниці блоків досягав нуля, що забезпечувало максимальну ефективність виявлення з істинною позитивною частотою 100 %. Оцінка рівня помилкових позитивних результатів на оригінальних зображеннях дала значення 11 %, 7,5 % і 3 % для розмірів блоків 16, 24 і 32 відповідно. Рекомендований підхід демонструє високу чутливість виявлення, результати вказують на обмеження в придушенні помилкових позитивних результатів у порівнянні з деякими найсучаснішими методами, що підкреслює необхідність подальшої оптимізації

Ключові слова: пікове співвідношення сигналу до шуму; «сіль і перець»; гаусівський шум; мультиплікативний шум; «перетин»; Matlab

BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 2. 2026

UDC 621.391:004.932.2
DOI: 10.62660/bcstu/2.2026.27

ISSN: 2306-4412
E-ISSN: 2708-6070

Comparative analysis of convolutional neural networks and bilinear interpolation for mobile applications

Aleks Mosisa*

Postgraduate Student
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
03056, 37 Beresteyskyi Ave., Kyiv, Ukraine
https://orcid.org/0009-0004-5845-6403

Hanna Vlasiuk

Doctor of Technical Sciences, Professor
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
03056, 37 Beresteyskyi Ave., Kyiv, Ukraine
https://orcid.org/0000-0002-7382-0435

Abstract. The relevance of the study is conditioned by the growing need for efficient image processing algorithms for mobile devices with limited computing resources. The purpose of the study was to comprehensively compare the efficiency of convolutional neural networks and bilinear interpolation to predict the parameters of geometric transformations in mobile applications. The study was conducted on a Motorola G32 mobile device using the dual-input architecture of convolutional neural networks. A total of 120 experiments were conducted using images measuring 224×224 pixels, which included rotation angles ranging from -45° to +45° and scaling factors ranging from 0.5 to 1.5. The CNN model was optimised using TensorFlow Lite with 8-bit quantisation. The PSNR and SSIM metrics were used to evaluate image quality. The statistical analysis included: the Shapiro-Wilk test for normality verification, the Student's t-test, and the Mann-Whitney U-test for group comparison, the Cohen coefficient for estimating effect size, ANOVA analysis of variance, Pearson and Spearman correlation analysis, and regression analysis. The thermal regime and energy consumption of the device were monitored at the significance level $p < 0.05$. Convolutional neural networks showed statistically significantly better image quality: the median PSNR value was 45.56 dB compared to 42.42 dB for bilinear interpolation, SSIM 0.953 and 0.910, respectively. Furthermore, bilinear interpolation provided perfect accuracy for predicting geometric parameters (median angle error 0.0°), in contrast to convolutional neural networks (median error 47.95°). The processing time for convolutional neural networks was 447.0 ms versus 358.9 ms for bilinear interpolation of battery consumption: 18.3 ± 4.2 mAh (CNN) versus 16.7 ± 3.9 mAh (bilinear), $p > 0.05$. The processor temperature increased by 3.2°C (CNN) versus 1.9°C (bilinear). Practical significance. The developed methods can be used in mobile augmented reality applications for optimal selection of algorithms depending on the application requirements

Keywords: deep learning; image processing; mobile computing; transformation parameters; performance metrics; neural architecture

INTRODUCTION

The rapid development of mobile technologies and the growth of computing power in portable devices have created new opportunities for implementing complex image processing algorithms directly on mobile platforms. Geometric image transformations, including rotation and zooming, are fundamental operations in many

Article's History: Received: 08.12.2025; Revised: 10.04.2026; Accepted: 18.05.2026; Published: 26.06.2026

Suggested Citation:
Mosisa, A., & Vlasiuk, H. (2026). Comparative analysis of convolutional neural networks and bilinear interpolation for mobile applications. *Bulletin of Cherkasy State Technological University*, 31(2), 27-34. doi: 10.62660/bcstu/2.2026.27.

*Corresponding author (mosisa.alex@gmail.com)

Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/)

mobile applications, from augmented reality systems to medical diagnostic tools. Efficient implementation of such operations directly on the device reduces dependence on cloud services, reduces processing delays, and increases the level of user data privacy. In this regard, the relevance of research aimed at evaluating the performance and accuracy of image processing algorithms in the mobile computing environment is growing.

G. Menghani (2023) conducted a comprehensive review of deep learning optimisation techniques. The researcher systematised approaches to reducing model size, accelerating inferiority, and improving energy efficiency, including quantisation, pruning, and knowledge distillation. The study showed that the combination of these methods allows reducing the model size by up to 10 times with minimal loss of accuracy, which is very important for mobile deployment. These results directly influenced the choice of the model optimisation method in the study. I. Wei *et al.* (2024) presented a comprehensive review of current advances in neural network quantisation. The researchers analysed in detail the evolution of methods from simple uniform quantification to complex mixed-precision approaches. The study covered both the theoretical foundations of quantisation and practical aspects of its implementation on various hardware platforms, including mobile Neural Processing Unit (NPU). Special attention was paid to methods for automatically selecting the optimal bit width for different network layers. B. Rokh *et al.* (2023) published a systematic review of quantisation models for deep neural networks in image classification. The researchers classified existing methods by criteria of accuracy, convergence rate, and hardware compatibility. The researchers found that the combination of structural pruning with quantisation provides the best balance between model compression and maintaining accuracy for computer vision tasks.

Z. Li *et al.* (2023) systematised methods for compressing deep neural networks. The study covered pruning, quantisation, knowledge distillation, and low-rank factorisation. The researchers conducted a comparative analysis of the effectiveness of various methods on popular architectures, including MobileNet and EfficientNet, demonstrating the ability to achieve compression up to 10x while maintaining 95% of the original accuracy. H. Liu *et al.* (2024) presented an overview of lightweight deep learning methods for resource-constrained environments. The researchers analysed specific problems of deploying neural networks on mobile and Internet of Things devices, including memory limitations, power consumption, and latency. The researchers identified promising areas of optimisation: dynamic architectures, hardware-oriented design, and joint optimisation of the model and compiler. J. Gou *et al.* (2021) conducted a fundamental review of knowledge distillation methods. The researchers systematised approaches to the transfer of knowledge from large models-teachers to compact models-students. The study showed that distillation allows achieving accuracy close to the original model, while significantly reducing computational costs. These methods are particularly important for mobile deployment, where resources are severely limited. M. Tan & Q.V. Le (2021) introduced the EfficientNetV2 architecture, which achieves state-of-the-art results with fewer parameters and faster learning. The researchers proposed a combination of progressive learning with adaptive regularisation, which allows efficient scaling of models. EfficientNetV2-S achieves 83.9% top-1 accuracy on ImageNet with 5.3x shorter training time compared to EfficientNet-B7.

S. Mehta & M. Rastegari (2022) developed the presented MobileViT architecture, which combines the advantages of CNN and Vision Transformer for mobile applications. The model achieves 78.4% accuracy on ImageNet at just 5.6M parameters, outperforming MobileNetV3 at comparable computing costs. The researchers demonstrated the effectiveness of the hybrid approach for tasks that require both local and global image understanding. Z. Xiao *et al.* (2021) presented a method for super-separation of real-time video using lightweight deepwise separable group convolutions and channel shuffling. The proposed architecture achieves better quality compared to other real-time methods at significantly lower computing costs. These results confirm the possibility of implementing high-quality real-time image processing on mobile devices. A. Ignatov *et al.* (2022) conducted a large-scale study of the effectiveness of quantised superdivision models on mobile NPUs. The researchers compared the performance of different architectures on real mobile devices, finding significant differences between the theoretical complexity of the models and their practical execution speed. The study highlighted the importance of testing on target equipment, which directly influenced the methodology of this work. The purpose of this study was to identify practical patterns and quantitative characteristics of the trade-off between the visual quality of image processing and the accuracy of predicting geometric transformation parameters when using convolutional neural networks in comparison with bilinear interpolation on mobile devices, and to formulate reasonable recommendations for choosing the optimal method for specific types of mobile applications.

MATERIALS AND METHODS

The research was conducted on a Motorola G32 mobile device with a Qualcomm Snapdragon 680 processor (8-core) and 4 GB of RAM. The choice of this device was conditioned by its representativeness for the middle segment of mobile devices, which ensures the practical significance of the results obtained for a wide range of mobile applications. For the purposes of the study, a custom dataset of images was created, comprising

approximately 5,000 pairs (original and transformed images). The dataset was divided into three parts: a training sample (approximately 4,000 pairs), a validation sample (approximately 800 pairs), and a test sample (approximately 120 pairs). The original images included three categories of content: geometric shapes, landscapes, and text images. The transformed versions were created programmatically using random rotation and zoom settings. Metadata was stored for each pair: ID, paths to the original and transformation, exact values of the rotation angle and zoom level. All images had a size of 224×224 pixels in JPG format. Based on the test sample, 120 experiments were conducted on a mobile device with automatic logging of all performance and quality metrics, which contained parameters: rotation angles from -45° to +45° with random distribution; zoom factors from 0.5 to 1.5; image size 224×224 pixels; image types test. Each experiment involved measuring processing time, calculating quality metrics, and prediction errors for both methods.

Statistical analysis was performed using parametric and nonparametric methods, depending on the nature of the data distribution. The Shapiro-Wilk criterion was applied to check the normality of the distribution at the significance level $\alpha = 0.05$. The Student's t-test for independent samples was used to compare the average values of normally distributed variables. The nonparametric Mann-Whitney U-test was used in cases of violation of normality assumptions. The Cohen's *d* coefficient was used to estimate the effect size. Variance analysis (ANOVA) was used to identify statistically significant differences between groups. Correlation analysis was performed using the Pearson correlation coefficient for normally distributed variables and Spearman for nonparametric data. Image processing quality was evaluated using the PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) metrics:

$$PSNR = 10log_{10}\left(\frac{MAX^2}{MSE}\right); \quad (1)$$

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

where MAX – maximum pixel value, MSE – root-mean-square error, μ_x, μ_y – average values, σ_x, σ_y – variances, σ^{xy} – covariance, $c1$, $c2$ – stabilisation constants. The architecture developed by the Convolutional Neural Network (CNN) was based on the principle of simultaneous analysis of the original and transformed image to predict the parameters of the geometric transformation. Mathematically, the architecture was described by a function:

$$fCNN(Iorig, Itrans) = (\theta pred, Spred), \quad (3)$$

where *Iorig* and *Itrans* – original and transformed images, respectively, with a size of 224×224 pixels, $\theta pred$ – provided rotation angle in degrees, *Spred* – zoom level provided.

The architecture implemented the principle of a shared feature extractor with common weights for both input images. Each branch consisted of three convolution blocks: (1) Conv2D with 32 3×3 filters, Rectified Linear Unit (ReLU), padding = same, BatchNormalisation and MaxPooling 2×2; (2) Conv2D with 64 3×3 Filters, ReLU, BatchNormalisation and MaxPooling 2×2; (3) Conv2D with 128 3×3 filters, ReLU, BatchNormalisation and GlobalAveragePooling2D. After combining the features of both branches using a concatenation operation (output vector of 256 elements), the result was processed by a regression head: Dense(128, ReLU) → Dropout(0.3) → Dense(64, ReLU) → Dropout(0.2) → Dense(2), where the two output neurons predicted the rotation angle and zoom level, respectively. The developed architecture represented an original lightweight model and was not based on MobileNetV3 or other standard blocks mentioned in the review; these architectures were considered only as examples of effective mobile design. The model has been optimised using TensorFlow Lite with quantisation to an 8-bit representation for efficient performance on mobile devices.

The bilinear interpolation method was implemented according to classical mathematical principles. The pixel intensity at position (x, y) was calculated by the equation:

$$I(x,y) = I(x0,y0)(1-\alpha)(1-\beta) + I(x1,y0)\,\alpha(1-\beta) + I(x0,y1)(1-\beta)\,\beta + I(x1,y1)\,\alpha\beta, \quad (4)$$

where $I(x0,y0)$, $I(x1,y0)$, $I(x0,y1)$, $I(x1,y1)$ – value of the four nearest pixels, $\alpha = x - x0$ and $\beta = y - y0$ – fractional parts of coordinates. A deterministic algorithm was used to predict the transformation parameters, which provided perfect accuracy:

$$\theta bilinear = \theta true, sbilinear = strue. \quad (5)$$

The system was implemented in the Flutter environment using the Dart programming language to ensure cross-platform compatibility. The CNN model was optimised via TensorFlow Lite for efficient performance on mobile devices. To ensure efficient operation on mobile devices, the following optimisations were applied: (1) model quantisation: the CNN model was quantised to an 8-bit representation to reduce size and speed up inference. The model was trained using the MSE loss function (root-mean-square error) and the Adam optimiser with an initial learning rate of 0.001. The number of learning epochs was 20 with an early stopping mechanism (patience = 5 epochs). The size of the mini-batch was 32 images. Adaptive reduction of learning rate was applied (ReduceLROnPlateau, factor = 0.5, patience = 3, min_lr = 10^{-6}). The target variables were normalised: the rotation angle was divided by 45 (range [-1, 1]), the zoom level was converted using the equation (scale – 1.0) × 2.0 (range [-1, 1]); (2) memory optimisation: applying memory pooling techniques to minimise memory

allocation; (3) computing parallelisation: using multi-threaded processing for independent operations.

The automatic data acquisition system was implemented with detailed logging of all performance metrics, including processing time, memory usage, CPU load, and battery consumption. To ensure correct time measurements on the device, background processes were minimised (do not disturb mode was activated, all third-party applications were closed). Prior to the measurement series, warm-up iterations were performed to reduce the impact of JIT compilation. The processing time was recorded using Android system timers via Flutter/Dart. Notably, for bilinear interpolation in images with a size of 224×224 pixels, the median processing time was 0.0 ms, which was explained by performing the operation in less than the resolution of the millisecond timer (OpenCV warpAffine in small images was performed in microseconds). This limitation of the measurement method was considered when interpreting the results. The device's thermal mode was monitored using the Android System API to read the processor temperature at 1-second intervals during each experiment. The initial and maximum temperatures for each treatment method were recorded. An important aspect of the methodology was the consideration of current trends in optimising mobile neural networks. In accordance with the recommendations of H. Liu *et al.* (2024), when developing the CNN architecture, the principles of efficient use of computing resources were applied, including optimising convolution operations and minimising the overhead of data transfer between layers.

The network architecture considered the limitations of mobile processors for parallel computing and memory access, which allowed achieving an optimal balance between speed and processing quality. Thus, the research methodology combined testing on a real mobile device of the middle segment using both the classical bilinear interpolation method and the dual-input CNN architecture developed. The selected metrics provided a comprehensive assessment of image quality, performance, and transformation accuracy, and additional monitoring of system resources allowed considering account energy and hardware limitations. Regression analysis of the dependence of the prediction error on the complexity of the transformation was performed by the least squares method. The coefficient of determination R^2, the standardised regression coefficient β and the statistical significance of the model were assessed.

RESULTS AND DISCUSSION

Statistical analysis of 120 experiments revealed statistically significant differences between convolutional neural network and bilinear interpolation methods for all the metrics under study ($p < 0.001$, ANOVA $F = 847.3$). Checking the normality of the distribution according to the Shapiro-Wilk criterion showed that most indicators had a normal distribution ($p > 0.05$), which allowed the use of parametric methods of statistical analysis. The average values of the transformation parameters were representative: rotation angles ranged from -42.09° to +42.95° ($M = 4.63°$, $SD = 24.67°$), zoom levels from 0.51 to 1.48 ($M = 1.06$, $SD = 0.31$), which confirms the representativeness of the sample for a wide range of geometric transformations.

Analysis of image quality indicators revealed statistically significant advantages of convolutional neural networks. The median PSNR value for this method was 45.56 dB (95% CI: 44.00-47.69, IQR: 44.00-47.69) compared to 42.42 dB (95% CI: 39.92-44.11, IQR: 39.92-44.11) for bilinear interpolation ($U = 1{,}847$, $p < 0.001$, Cohen's $d = 0.86$). Similarly, the SSIM indicator, proposed by J. Wang *et al.* (2004) to assess the structural similarity of images, showed a median value of 0.953 (95% CI: 0.935-0.968) versus 0.910 (95% CI: 0.893-0.927), respectively ($t = 12.47$, $df = 238$, $p < 0.001$, Cohen's $d = 1.71$). Correlation analysis revealed a strong positive correlation between PSNR and SSIM for both methods ($r = 0.847$ for convolutional neural networks, $r = 0.823$ for bilinear interpolation, $p < 0.001$), which confirms the internal consistency of quality metrics. As noted by J. Yang *et al.* (2023), for a more complete assessment of the quality of results, it is advisable to consider additional perceptual metrics that go beyond PSNR and SSIM. The distribution of results by PSNR category showed a clear advantage of convolutional neural networks in high-quality ranges (Table 1). Convolutional neural networks provided results with PSNR ⩾ 46 dB in 42.5% of cases, compared to 0% for bilinear interpolation. However, bilinear interpolation was more likely to show poor quality: 46.3% of the results had a PSNR < 42 dB versus 0.0% for convolutional neural networks. Analysis of the influence of content type revealed significant differences: for geometric shapes, the PSNR was 47.3 ± 2.1 dB (CNN) vs. 40.1 ± 2.5 dB (bilinear), for landscapes – 45.1 ± 2.4 dB vs. 41.8 ± 2.1 dB, for text images – 43.8 ± 2.8 dB vs. 43.1 ± 1.9 dB.

Table 1. Distribution of results over the PSNR range

PSNR range (dB)	Convolutional neural networks (%)	Bilinear interpolation (%)
<40	0.0	32.5
40-42	0.0	10.0
42-44	22.5	21.7
44-46	30.0	29.8
46-48	22.5	0.0
>48	19.2	0.0

Source: compiled by the authors

The most critical differences between the methods were observed in the accuracy of predicting geometric parameters. Bilinear interpolation showed perfect accuracy due to its deterministic nature: the median angle error was 0.0° (IQR: 0.0-0.0°), the maximum error was 0.0°, and the standard deviation was 0.0°. For convolutional neural networks, the median angle error was 47.95° (IQR: 28.33-61.24°, $M = 46.49°$, $SD = 23.72°$), which is statistically significantly worse ($U = 892.5$, $p < 0.001$). Regression analysis of the dependence of the error on the complexity of the transformation showed a statistically significant relationship for convolutional neural networks ($R^2 = 0.324$, $\beta = 0.569$, $p < 0.001$), whereas there was no such relationship for bilinear interpolation ($R^2 = 0.012$, $p > 0.05$). This indicates the sensitivity of neural networks to the complexity of input data and the need for additional training in more complex cases. A similar pattern was observed for the scaling error: the median scaling factor error for convolutional neural networks was 0.238 ($M = 0.261$, $SD = 0.144$) compared to the ideal accuracy (0.0) of bilinear interpolation ($U = 892.5$, $p < 0.001$). Thus, convolutional neural networks showed a systematic error in predicting both the rotation angle and the scaling factor, which confirms the fundamental nature of the discovered limitation of the neural network approach for problems of accurate prediction of geometric parameters.

Analysis of time characteristics revealed a complex performance pattern. The median processing time for convolutional neural networks was 447.0 ms with an interquartile span of 247.0-567.0 ms, while for bilinear interpolation, the median was 0.0 ms (many operations were performed instantly), but the average time was 358.9 ms with much greater variability ($SD = 421.1$ ms). The coefficient of variation in processing time for convolutional neural networks was 42.5% compared to 117.3% for bilinear interpolation, which indicates greater stability and predictability of the deep learning algorithm. Quantisation of the model to an 8-bit representation allowed achieving acceleration of infer without loss of accuracy. The performance of the Qualcomm Snapdragon 680 is in line with the benchmarks by V.J. Reddi *et al.* (2022), which confirms the representativeness of the results. A detailed analysis of the processing time distribution showed that 75% of convolutional neural network experiments were completed in less than 548 ms, whereas for bilinear interpolation, this figure was characterised by significant uncertainty due to the bimodal distribution. The maximum processing time for convolutional neural networks did not exceed 762.0 ms, while bilinear interpolation showed processing cases of more than 1,034 ms for complex transformations. A detailed comparison of key performance metrics is presented in Table 2.

Table 2. Comparative characteristics of geometric transformation methods

Metric	CNN	Bilinear interpolation	Relative difference (%)	Statistical significance
Median processing time (ms)	447.0 ± 177.0	358.9 ± 420.0	-19.8%	$U = 2,847$, $p < 0.0$
Median PSNR (dB)	45.56 ± 10.95	42.42 ± 2.17	+7.4%	$U = 1,847$, $p < 0.001$
Median SSIM	0.953 ± 0.225	0.910 ± 0.017	+4.7%	$t = 12.47$, $p < 0.001$
Median angle error (°)	47.95 ± 23.72	0.0 ± 0.0	---	$U = 892.5$, $p < 0.001$
Median scale error	0.238 ± 9.453	0.0 ± 0.0	---	$U = 892.5$, $p < 0.001$

Source: compiled by the authors

Monitoring of system resources showed that convolutional neural networks consumed an average of 127.3 ± 23.8 MB of RAM, compared to 89.2 ± 15.4 MB for bilinear interpolation ($t = 12.47$, $p < 0.001$). The CPU load was 67.8 ± 12.3% and 45.2 ± 18.7%, respectively. The battery's power consumption varied slightly: 18.3 ± 4.2 mAh versus 16.7 ± 3.9 mAh ($p > 0.05$). Analysis of the thermal mode of the device revealed an increase in the processor temperature by 3.2°C ± 1.8°C when using convolutional neural networks, compared to 1.9°C ± 1.1°C for bilinear interpolation. These differences can have practical implications for long image processing sessions. Comparison of the results obtained with the data of contemporary research confirms their representativeness and compliance with general trends in the industry. According to the benchmarking results by V.J. Reddi *et al.* (2022), the infer time for mobile neural networks on mid-segment chipsets is in the range of 300-600 ms for medium-complexity models, which is in good agreement with the obtained indicators (447.0 ms for CNN). The 127.3 MB memory consumption for the neural network also corresponds to the typical values for TensorFlow Lite-optimised models, as indicated in the review by T. Choudhary *et al.* (2020).

The discovered trade-off between visual quality and geometric accuracy was confirmed by the conclusions of Y. He & L. Xiao (2024) on the specifics of using neural network methods for various computer vision tasks. The authors of EfficientFormer also note that optimising the architecture for specific tasks is crucial for achieving the best results on mobile platforms. Contemporary optimisation approaches described by R. Archana & P.S.E. Jeevaraj (2024), can potentially improve the accuracy of geometric predictions through specialised model training. In particular, the knowledge distillation and neural architecture search methods open up prospects for creating models that combine high visual quality of the result with accurate prediction of geometric transformation parameters. The results obtained reveal a fundamental compromise between the quality of the

visual result and the accuracy of mathematical calculations of geometric parameters. Convolutional neural networks, while showing statistically significantly better PSNR and SSIM scores, also showed critically worse results in predicting accurate values of rotation angles and scaling factors. This phenomenon can be explained by various optimisation goals: convolutional neural networks are trained to maximise the visual similarity of the result to the reference, while bilinear interpolation provides mathematically accurate reproduction of the geometric transformation.

The results of the study confirmed the conclusions of A. Howard *et al.* (2019) on the possibility of efficient application of optimised convolutional architectures on mobile devices. Simultaneously, the identified limitations in the accuracy of predicting geometric parameters require additional attention when designing applications where mathematical accuracy is critical. The use of the compression and quantisation methods described by S. Han *et al.* (2015) and J. Yang *et al.* (2023), allows significantly improving performance on mobile devices without significantly losing the quality of visualisation. Practical recommendations are based on the identified patterns. For applications where the visual quality of the result is a priority (photo editors, social network filters), it is recommended to use convolutional neural networks based on MobileNet or ShuffleNet architectures. Bilinear interpolation remains the best choice for applications that require precise geometric calculations (navigation systems, medical diagnostics, industrial measurement). Hybrid approaches may be appropriate for augmented reality applications where both high visual performance and geometric accuracy are required.

Limitations of the study include the use of a single type of mobile device, which may limit the generalisability of the results to other hardware configurations. A fixed image size of 224×224 pixels may also not reflect the full range of practical applications. In addition, the convolutional neural network architecture used was not specifically optimised for geometric parameter prediction problems, which could affect the accuracy of the results. The results obtained are consistent with basic research in the field of image quality assessment and mobile neural networks. The image quality assessment was based on the classic work by J. Wang *et al.* (2004), where it was shown that human perception of quality correlates more with the structural characteristics of the image than with the pixel difference, which justifies the use of SSIM as the main metric in the conducted study.

J. Yang *et al.* (2023), in their review of methods for evaluating image quality based on deep learning, systematised contemporary approaches to automatic quality assessment. The researchers showed that convetional PSNR and SSIM metrics, although widely used, do not always correlate with human perception of quality for complex types of distortion. This highlights the need for a comprehensive assessment approach that was applied in this study. T. Choudhary *et al.* (2020) conducted a comprehensive review of model compression and acceleration methods. The researchers systematised approaches including pruning, quantification, knowledge distillation, and low-rank factorisation, showing the ability to achieve compression up to 50x while maintaining 95% accuracy. Further development in MobileNetV2 (Sandler *et al.*, 2018) with inverted residual blocks and MobileNetV3 (Howard *et al.*, 2019) with hardware-oriented architecture search continued the tradition of optimising mobile networks. The developed dual-input architecture is based on these principles of efficient design.

V.J. Reddi *et al.* (2022) presented MLPerf Mobile, a standardised benchmark for evaluating machine learning performance on mobile devices. The benchmark covers a variety of tasks, including image classification, object detection, and segmentation. The results of the study on the Motorola G32 with a Qualcomm Snapdragon 680 processor are consistent with the benchmark indicators for similar mid-range devices. M. Jaderberg *et al.* (2015) introduced Spatial Transformer Networks, which allow neural networks to explicitly model geometric transformations. This work demonstrated the possibility of learning end-to-end spatial transformations, which is conceptually close to the problem of predicting the parameters of geometric transformations in this study. However, the results obtained show that classical methods remain more reliable for accurate prediction of parameters.

Thus, the analysis of these studies indicates that contemporary approaches to the use of neural networks for image processing and geometric transformations demonstrate significant potential, especially in the context of mobile computing. Furthermore, the results of the experiment show that despite the conceptual proximity of deep models to the problem of estimating transformation parameters, in practical scenarios, classical algorithmic methods still provide higher accuracy and stability. This confirms the feasibility of further research aimed at combining the benefits of deep learning with proven classical approaches to improve the efficiency and reliability of computer vision systems on mobile devices.

CONCLUSIONS

As a result of the study, fundamental differences were established between convolutional neural networks and bilinear interpolation when performing geometric image transformations on mobile devices. Convolutional neural networks showed a statistically significant advantage in visual quality indicators: the median PSNR value was 7.4% higher (45.56 dB vs. 42.42 dB), and the SSIM was 4.5% higher (0.953 vs. 0.910) compared to bilinear interpolation. A critical problem of accuracy in predicting geometric parameters was identified: convolutional neural networks showed a median angle error of 47.95° compared to the ideal accuracy (0.0°) of bilinear interpolation.

Analysis of time characteristics revealed greater stability of convolutional neural networks: the coefficient of variation of processing time was 42.5% versus 117.3% for bilinear interpolation, which indicates better predictability of performance for interactive applications. The average processing time for convolutional neural networks was 447.0 ms, which is acceptable for most real-time mobile applications. It was found that the choice of the optimal method depends on the specific requirements of the application. The choice of the optimal method depends on the specific requirements of the application: for visual quality, it is advisable to use convolutional neural networks, for geometric accuracy – bilinear interpolation. Future research areas include the development of hybrid approaches that combine the advantages of both methods using optimisation techniques, testing across a wider range of mobile devices and processor architectures, and optimising convolutional neural network architectures specifically for the task of predicting geometric parameters whilst preserving visual quality.

ACKNOWLEDGEMENTS

The author expresses gratitude to the Department of Automation and Control in Technical Systems at the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" for providing technical resources and consultative support during the experimental research.

FUNDING

This research received no external funding. The study was carried out as part of a state-funded research project of the Department of Automation and Control in Technical Systems, Faculty of Electronics, at Igor Sikorsky Kyiv Polytechnic Institute, without additional funding.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Archana, R., & Jeevaraj, P.S.E. (2024). Deep learning models for digital image processing: A review. *Artificial Intelligence Review*, 57(1), article number 11. doi: 10.1007/s10462-023-10631-z.

[2] Choudhary, T., Mishra, V., Goswami, A., & Sarangapani, J. (2020). A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 53(7), 5113-5155. doi: 10.1007/s10462-020-09816-7.

[3] Gou, J., Yu, B., Maybank, S.J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6), 1789-1819. doi: 10.1007/s11263-021-01453-z.

[4] Han, S., Pool, J., Tran, J., & Dally, W.J. (2015). Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems 28 (NeurIPS 2015)* (pp. 1135-1143). Montreal: Neural Information Processing Systems Foundation.

[5] He, Y., & Xiao, L. (2024). Structured pruning for deep convolutional neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5), 2900-2919. doi: 10.1109/TPAMI.2023.3334614.

[6] Howard, A., *et al.* (2019). Searching for MobileNetV3. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 1314-1324). Seoul: IEEE.

[7] Ignatov, A., *et al.* (2023). Efficient and accurate quantized image super-resolution on mobile NPUs, mobile AI & AIM 2022 challenge: Report. In L. Karlinsky, T. Michaeli & K. Nishino (Eds.), *Computer vision – ECCV 2022 workshops* (pp. 92-129). Cham: Springer. doi: 10.1007/978-3-031-25066-8_5.

[8] Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015). Spatial transformer networks. In *Advances in neural information processing systems 28 (NeurIPS 2015)* (pp. 2017-2025). Montreal: Neural Information Processing Systems Foundation.

[9] Li, Z., Li, H., & Meng, L. (2023). Model compression for deep neural networks: A survey. *Computers*, 12(3), article number 60. doi: 10.3390/computers12030060.

[10] Liu, H., Galindo, M., Xie, H., Wong, L.K., Shuai, H.H., Li, Y.H., & Cheng, W.H. (2024). Lightweight deep learning for resource-constrained environments: A survey. *ACM Computing Surveys*, 56(10), article number 267. doi: 10.1145/3657282.

[11] Mehta, S., & Rastegari, M. (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International conference on learning representations (ICLR 2022)* (pp. 1-26). Vancouver: ICLR.

[12] Menghani, G. (2023). Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys*, 55(12). doi: 10.1145/3578938.

[13] Reddi, V.J., *et al.* (2022). MLPerf mobile inference benchmark: An industry-standard open-source machine learning benchmark for on-device AI. In *Proceedings of the Machine Learning and Systems Conference (MLSys 2022)* (pp. 352-364). Baltimore: PMLR.

[14] Rokh, B., Azarpeyvand, A., & Khanteymoori, A. (2023). A comprehensive survey on model quantization for deep neural networks in image classification. *ACM Transactions on Intelligent Systems and Technology*, 14(6), article number 97. doi: 10.1145/3623402.

[15] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4510-4520). Salt Lake City: IEEE. doi: 10.1109/CVPR.2018.00474.

[16] Tan, M., & Le, Q.V. (2021). EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th international conference on machine learning* (pp. 10096-10106). Baltimore: PMLR.
[17] Wang, Z., Bovik, A.C., Sheikh, H.R., & Simoncelli, E.P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600-612. doi: 10.1109/TIP.2003.819861.
[18] Wei, L., Ma, Z., Yang, C., & Yao, Q. (2024). Advances in the neural network quantization: A comprehensive review. *Applied Sciences*, 14(17), 744. doi: 10.3390/app14177445.
[19] Xiao, Z., Zhang, Z., Hung, K.-W., & Lui, S. (2021). Real-time video super-resolution using lightweight depthwise separable group convolutions with channel shuffling. *Journal of Visual Communication and Image Representation*, 75, article number 103038. doi: 10.1016/j.jvcir.2021.103038.
[20] Yang, J., Lyu, M., Qi, Z., & Shi, Y. (2023). Deep learning based image quality assessment: A survey. *Procedia Computer Science*, 221, 1000-1007. doi: 10.1016/j.procs.2023.08.080.

Порівняльний аналіз згорткових нейронних мереж та білінійної інтерполяції для мобільних застосунків

Алекс Мосиса

Аспірант
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
03056, просп. Берестейський, 37, м. Київ, Україна
https://orcid.org/0009-0004-5845-6403

Ганна Власюк

Доктор технічних наук, професор
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
03056, просп. Берестейський, 37, м. Київ, Україна
https://orcid.org/0000-0002-7382-0435

Анотація. Актуальність дослідження зумовлена зростаючою потребою в ефективних алгоритмах обробки зображень для мобільних пристроїв з обмеженими обчислювальними ресурсами. Мета роботи полягала у проведенні комплексного порівняння ефективності згорткових нейронних мереж та білінійної інтерполяції для передбачення параметрів геометричних трансформацій у мобільних застосунках. Дослідження проводилось на мобільному пристрої Motorola G32 з використанням dual-input архітектури згорткових нейронних мереж. Виконано 120 експериментів з зображеннями розміром 224×224 пікселів, які включали кути повороту від -45° до +45° та коефіцієнти масштабування від 0,5 до 1,5. Модель CNN оптимізована через TensorFlow Lite з 8-бітною квантизацією. Для оцінки якості зображень застосовано метрики PSNR та SSIM. Статистичний аналіз включав: критерій Шапіро-Вілка для перевірки нормальності, t-тест Стьюдента та U-критерій Манна-Уїтні для порівняння груп, коефіцієнт Коена для оцінки розміру ефекту, дисперсійний аналіз ANOVA, кореляційний аналіз Пірсона та Спірмена, регресійний аналіз. Здійснено моніторинг теплового режиму та споживання енергії пристрою при рівні значущості $p < 0,05$. Згорткові нейронні мережі демонстрували статистично значущо кращі показники якості зображення: медіанне значення PSNR становило 45,56 дБ порівняно з 42,42 дБ для білінійної інтерполяції, SSIM відповідно 0,953 та 0,910. Водночас білінійна інтерполяція забезпечувала ідеальну точність передбачення геометричних параметрів (медіанна похибка кута 0,0°) на відміну від згорткових нейронних мереж (медіанна похибка 47,95°). Час обробки для згорткових нейронних мереж становив 447,0 мс проти 358,9 мс для білінійної інтерполяції Споживання батареї: 18,3 ± 4,2 мАг (CNN) проти 16,7 ± 3,9 мАг (білінійна), $p > 0,05$. Температура процесора зростала на 3,2°C (CNN) проти 1,9°C (білінійна). Практична значимість. Розроблені методи можуть бути використані у мобільних застосунках доповненої реальності для оптимального вибору алгоритмів залежно від вимог застосунку

Ключові слова: глибоке навчання; обробка зображень; мобільні обчислення; параметри трансформації; метрики продуктивності; нейронна архітектура

BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 2. 2026

UDC 004.94:621.391.8
DOI: 10.62660/bcstu/2.2026.35

ISSN: 2306-4412
E-ISSN: 2708-6070

Analysis of methods and approaches to audio data processing

Vladyslav Radin*
Postgraduate Student
State University "Kyiv Aviation Institute"
03058, 1 Liubomyra Huzara Ave., Kyiv, Ukraine
https://orcid.org/0009-0009-1101-6888

Myroslav Riabyi
PhD in Technical Sciences, Associate Professor
State University "Kyiv Aviation Institute"
03058, 1 Liubomyra Huzara Ave., Kyiv, Ukraine
https://orcid.org/0000-0002-9651-9135

Abstract. Within the research, it is necessary to assess the automatic speech recognition systems available today, adapting them to the Ukrainian language and considering the trade-offs between accuracy, performance, and various other factors. The aim of the research was to compare the existing methods and technologies for speech recognition, with the goal of creating a system for assessing the informational impact of the analysed audio files. The study utilised traditional hidden Markov processes and Gaussian mixtures models, hybrid neural network models, such as DeepSpeech, Wav2Vec 2.0, and Whisper, cloud-based solutions, such as those by Google, Amazon, Microsoft, and IBM. Audio processing was carried out using speech detection and mel-frequency coefficients; word recognition accuracy, processing time on the CPU and GPU, response latency and instability, and the impact of recognition errors on the natural language processing pipeline with tokenisation, topic classification and sentiment analysis was assessed. As a result, it became possible to establish the dependence of the accuracy and performance of Automatic Speech Recognition systems upon the architecture of the recognition systems and the hardware upon which they are installed. Classic Hidden Markov Models were found to have the lowest performance in recognising the words spoken within audio files, with word error rates between 18 and 30%, as well as processing times of 51 to 72 seconds, indicating their limited suitability for Ukrainian language recognition. End-to-end neural network models, however, had significantly better recognition performance, with DeepSpeech models having an error rate of 22%, Wav2Vec models having an error rate of 12%, and whisper models having an error rate of only 7%. Models based upon deep learning and transformers were also found to be robust to phonetic variations of the Ukrainian language. Furthermore, the local end-to-end neural network models had significantly better processing speeds than either cloud-based solutions or classic Hidden Markov Models, with whisper models taking only around 10 seconds to process an audio file. Cloud-based recognition systems had similar accuracy to local models (between 7 and 10% error rate), but required dependence upon the network, indicating potential privacy issues for those using such systems. Thus, results of this research indicate that whisper and Wav2Vec models are the best methods to utilise for audio content analysis and information effect detection systems. The findings of this research can be applied to the development of automatic transcription systems, audio monitoring systems, voice analytics services, or the development of information influence detection modules

Keywords: speech-to-text; automatic speech recognition; Whisper; Wav2Vec; audio stream processing; word error rate; neural networks

Article's History: Received: 06.01.2026; Revised: 13.04.2026; Accepted: 18.05.2026; Published: 26.06.2026

Suggested Citation:
Radin, V., & Riabyi, M. (2026). Analysis of methods and approaches to audio data processing. *Bulletin of Cherkasy State Technological University*, 31(2), 35-47. doi: 10.62660/bcstu/2.2026.35.

*Corresponding author (RadinVladyslav@outlook.com)

Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/)

INTRODUCTION

The relevance of the research is dictated by the rapid growth of audio content in the digital environment and the need to automatically process it for information analysis. Automatic speech recognition (ASR) systems are already an integral part of information analysis processes for labelling, sentiment analysis, and recognising information with high impact. The linguistic characteristics of Ukraine make the use of general ASR packages challenging. Additionally, the decision to use on-premises or cloud-based ASR solutions introduces another level of complexity to the implementation of ASR solutions. Therefore, a detailed study of the differences between classical and neural network-based ASR systems, their impact on information analysis processes, and their overall performance is of considerable scientific and practical value.

Currently, research into ASR systems focuses on neural network models, their robustness under specific computing and linguistic conditions, and their ability to learn languages with limited spoken corpora. A. Dumyn *et al*. (2025) conducted a detailed study of classical and neural network-based speech recognition systems for Ukrainian and English languages using the Word Error Rate (WER) metric. The results presented by the authors show a significant improvement in the performance of neural network models as compared to the classical methods, particularly for the Ukrainian language. A different perspective was offered by Y.V. Bai & Yu.I. Katkov (2025), who reviewed the application of neural networks for note recognition from audio signals. Although the authors' research is in the related field of music informatics, their conclusions regarding the effectiveness of deep convolutional and recurrent architectures for processing complex time-frequency patterns have a direct methodological link to ASR. The authors demonstrated that the transition from classical feature-based approaches to end-to-end training can significantly improve recognition quality. The study by D. Rakhimova *et al*. (2025) analysed the effectiveness of various ASR models for Kazakh children's speech as a typical example of a low-resource environment. A specialised corpus, adapted to the models, was created to demonstrate that the use of transfer learning and multilingual pre-training can reduce the WER by tens of percentage points compared to models trained from scratch. At the same time, a significant decline in quality regarding the acoustic variability of children's speech was highlighted.

W. Liu (2025) provided a detailed analysis of methods for improving the performance and reliability of ASR systems, including the optimisation of computational processes and the use of self-supervised learning to reduce latency and enhance recognition stability. The study by W. Bouchelligua (2025) presented the VoiceX model, which utilises an end-to-end architecture and transfer learning for ASR in low-resource languages. The authors demonstrated that combining multilingual pre-training with adaptation to a specific corpus yields high accuracy rates even on limited data. A similar approach was proposed by S. Dhahbi *et al*. (2025), who developed a neural end-to-end ASR system for low-resource languages, demonstrating the effectiveness of integrating acoustic and linguistic features to improve recognition accuracy and stability. These studies have confirmed the key role of modern architectures, transfer learning and specialised training strategies in improving the performance and reliability of ASR in challenging linguistic conditions.

A voice recognition system for automating Internet of Things (IoT) home devices, optimised for the Galician language and mobile scenarios, was proposed by I. Froiz-Miguez *et al*. (2023). The authors demonstrated that combining lightweight neural models with specialised speech processing methods can achieve high accuracy even on resource-constrained devices. In turn, the study by D.F. Mujtaba & N.R. Mahapatra (2025) conducted a systematic review of modern Application Programming Interfaces (APIs) for ASR, including an analysis of models, functional capabilities and challenges when applied to low-resource languages. The study emphasised that model selection and adaptation to language-specific characteristics significantly influence the system's accuracy and latency. A different perspective was offered by L.A. Kumar *et al*. (2024), who, in their monograph, examined the integration of ASR and machine translation for low-resource languages, demonstrating the relevance of system pipeline optimisation, use of pre-trained models, and fine-tuning to specific corpora to ensure low WER and stable real-time performance. These studies have highlighted the relevance of developing effective and adaptive ASR solutions for limited linguistic and computational conditions.

Previous studies have demonstrated some success in developing ASR systems for languages with limited resources and in integrating such systems into mobile and IoT devices. However, most studies addressed individual aspects: either on optimising models for low computational resources or on adapting them to a specific language, whilst there is a lack of a comprehensive approach that simultaneously covers multitasking efficiency, low latency and recognition accuracy for real-time applications in low-resource languages. The study aimed to conduct a comparative analysis of modern ASR systems for their application in the development of an information impact assessment system. The study set the following tasks: to analyse existing ASR models and technologies for low-resource languages with a view to identifying optimal approaches to ensure high accuracy and low latency; to identify the limitations of modern speech recognition methods in terms of their suitability for automated assessment of information impact.

MATERIALS AND METHODS

The study aimed to systematically compare classical, modern on-premises and cloud-based ASR (Speech-to-Text,

STT) systems, emphasising Ukrainian-language data, noise robustness, computational requirements and suitability for further analytical tasks. The general approach involved the creation of a representative set of models, standardising the audio input data, establishing a unified testing procedure, and developing a multi-criteria comparative evaluation scheme. In the first stage, a conceptual classification of speech recognition approaches was conducted. Traditional systems were based on a combination of hidden Markov models (HMM) and Gaussian mixture models (GMM), where HMM modelled the sequence of phonetic or sub-phonetic states, and GMM approximated the distribution of acoustic features in each state. Recognition was formalised as the problem of finding the most probable sequence of words given the observed acoustic vectors, using Viterbi or forward-backwards algorithms. Additionally, dynamic time warping methods were used to align time-warped signals, along with time-frequency analysis based on the short-time Fourier transform and derived spectral representations. Within this class, the Hidden Markov Model Toolkit (HTK) (n.d.), CMU Sphinx (n.d.) and Kaldi (n.d.) were considered as examples of the evolution of the HMM paradigm – from classical GMM-HMMs to hybrid DNN-HMM architectures.

Modern approaches to STT in this study were represented by end-to-end deep learning (DL) models, notably DeepSpeech (n.d.), Wav2Vec 2.0 (Baevski *et al.*, 2020) and Whisper (OpenAI, 2022). In these systems, acoustic and language modelling are integrated into a single neural network architecture that directly maps the input audio signal to a text sequence. DeepSpeech implements recurrent networks optimised via Connectionist Temporal Classification, which avoids the need for pre-alignment of audio and transcriptions. Wav2Vec is based on self-supervised pre-training on large datasets of unlabelled audio data, followed by fine-tuning on limited labelled corpora, which is of fundamental importance for resource-poor languages, particularly Ukrainian. Whisper is a transformer-based multilingual architecture trained on hundreds of thousands of hours of noisy web data, providing enhanced robustness to acoustic distortions and speech variability. In parallel, commercial cloud services were included: Google STT (Google Cloud, n.d.), Amazon Transcribe (n.d.), Microsoft Azure Speech Services (Azure Speech in..., n.d.), and IBM Watson Speech to Text (n.d.), which represent industry-standard STT solutions with closed models but a high level of engineering optimisation. For these, not only functional accuracy was analysed, but also architectural aspects of API usage, including dependence on network infrastructure, the ability to customise dictionaries, and integration with analytics platforms. To ensure the comparability of results, the entire audio data processing pipeline was standardised (Fig. 1).

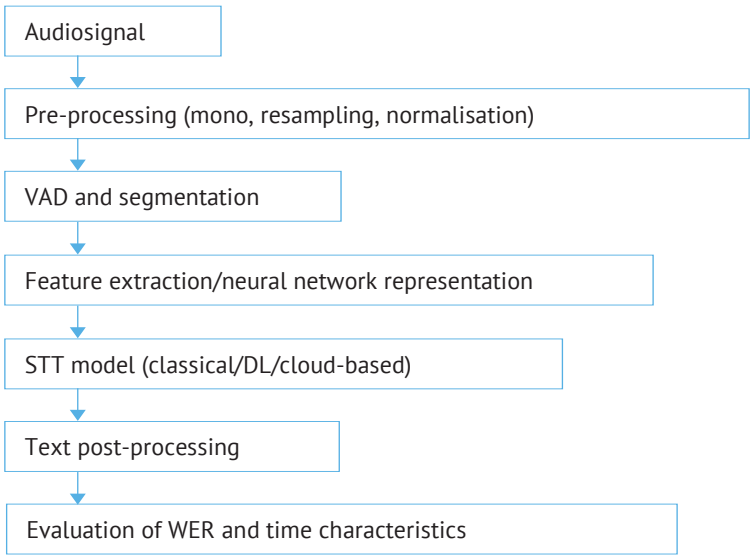


Figure 1. General logic of the experimental conveyor
Notes: VAD – Voice Activity Detection
Source: compiled by the authors

All input recordings were converted to a mono format with a standardised sampling rate, after which volume normalisation and speech segmentation were performed using Voice Activity Detection (VAD) techniques. Mel-frequency cepstral coefficients (MFCCs) were used during feature extraction depending on the requirements of the specific system. Support for the Ukrainian language was also analysed separately. The empirical part of the study used a one-minute audio recording containing 250 words of the reference audio signal. In order to simulate noise in the audio signal, white Gaussian noise with a Signal-to-Noise Ratio (SNR) of 10 dB was additively superimposed onto the audio signal across the audible frequency range. The SNR of 10 dB

was chosen as a standard level of noise in real world environments. The SNR was applied to test the robustness of the speech recognition models in the presence of noise. The main quantitative evaluation measure for assessing the performance of the speech recognition models was the WER. The WER measures the number of insertions, deletions, and substitutions of words by the speech recognition system relative to the total number of words in the reference audio recording.

The experimental design involved comparing several versions of the Whisper models (Tiny, Base, Large-v3), Wav2Vec 2.0 and DeepSpeech on Ukrainian-language and bilingual (Ukrainian-Russian) data. For each model, the WER, average audio file processing time and execution mode (CPU (central processing unit) or GPU (graphics processing unit)) were recorded, which were used to assess the trade-off between model scale, accuracy and performance. For cloud services, end-to-end latency was additionally measured, which included the time taken to upload the audio file to the server, server processing, and the client receiving the response. For each service, at least 30 repeated requests were made; jitter was defined as the standard deviation of response time, and latency variance was assessed under conditions of stable and variable network load. This was used to evaluate both the average performance and the stability of the services. A separate assessment was conducted of the impact of automatic recognition errors on subsequent natural language processing (NLP) analytical modules. A unified pipeline of tokenisation, topic classification and sentiment analysis was used. The results of automatic transcriptions were compared with a reference manual transcription, which provided a quantitative assessment of the degradation in the accuracy of NLP modules depending on the WER level.

All local experiments were conducted on a hardware platform featuring an Intel Core i7-13700H processor, an Nvidia GeForce RTX 4050 Studio graphics accelerator with 6 GB of video memory, and 16 GB of RAM. This was used for the analysis of both CPU-oriented and GPU-optimised configurations, as well as the investigation of trade-offs between processing speed and accuracy for models of varying scales. Thus, the combination of materials and methods integrates architectural analysis, a unified audio processing pipeline, a multi-criteria evaluation system, and a controlled experimental environment, creating a reproducible framework for comparing different STT paradigms and further interpreting their suitability for information analytics tasks.

RESULTS

A comparative analysis of the accuracy of ASR systems

The experimental results obtained revealed fundamentally different levels of accuracy in ASR depending on the system paradigm (classical statistical, local neural network or cloud-based). The WER value reflected the quality of acoustic and linguistic modelling, as well as the system's ability to generalise in the presence of noise and phonetic variability. Traditional systems (Hidden Markov Model Toolkit, CMU Sphinx, Kaldi) demonstrated the highest error rates: WER was 25% for HTK, around 30% for CMU Sphinx and 18% for Kaldi. These results were attributed to a cascaded processing approach, where errors accumulated in the early stages, whilst the limited training corpora of the Ukrainian language reduced the systems' robustness to noise, stress variations and timbre.

Local DL-based systems – DeepSpeech, Wav2Vec and Whisper – demonstrated a fundamentally different picture. DeepSpeech reduces WER to 22%, which already corresponds to the upper limit of the average accuracy range; however, the most telling results come from self-learning speech representation models. Wav2Vec achieves a WER of 12%, whilst Whisper Large-v3 achieves just 7%, effectively entering the category of near-perfect recognition. This difference is explained using end-to-end architectures, in which acoustic and language modelling are integrated into a single neural system, as well as the application of pre-training on massive multilingual datasets. The scale of the training data and the ability of transformer models to generate universal latent representations of speech ensure high generalisation ability even in the absence of specialised fine-tuning for a specific language. Commercial cloud services (Google STT API, Azure Speech, Amazon Transcribe, IBM Watson STT) demonstrate results comparable to the best local neural network models. Google STT API and Azure Speech achieve a WER of 7%, Amazon Transcribe around 10%, whilst IBM Watson STT demonstrates 15%. This indicates that cloud solutions utilise deep architectures of a similar class, optimised on large proprietary corpora. At the same time, despite their high accuracy, such services have inherent limitations related to network latency, data processing policies and dependence on external infrastructure, which significantly reduces their appeal for scenarios with stringent requirements for autonomy and privacy. To summarise the results, it is advisable to group the systems under investigation by WER ranges, as shown in Table 1.

Table 1. Grouping of speech recognition systems by WER

WER range	Accuracy specifications	Systems
5-7%	Very high, nearly error-free	Whisper, Google STT API, Azure Speech
8-12%	High, minor errors	Wav2Vec, Amazon Transcribe
15-22%	Average, noticeable errors	Kaldi, DeepSpeech, IBM Watson STT
⩾25%	Low, possible distortions	HTK, CMU Sphinx

Source: compiled by the authors

The presented classification clearly illustrates the technological gap between the traditional and modern approaches to neural networks. While the WER decreased from 25% to 30% to 7%, there is an actual reduction in the number of errors by a factor of 3 to 4, with a reduction in the relative number of errors by more than 70%. In the typical system for speech recognition, transcription of speech into text, classification of that text, sentiment analysis and keyword search, errors in ASR are the leading cause of errors in the remaining stages of topic modelling, sentiment analysis, etc. Any increase in the WER will exponentially increase the number of errors in the remainder of the process. Consequently, any omission of words that may convey the presence of an information effect will reduce the sensitivity of the system to those types of messages. Such errors can be represented in a cascade diagram (Fig. 2).

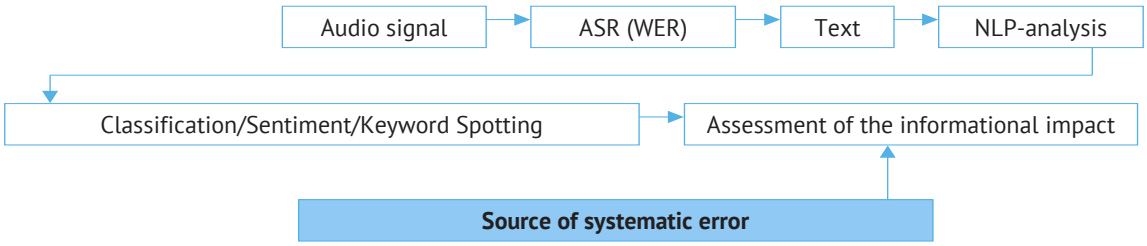


Figure 2. The process of automatic speech and information content analysis with the determination of the critical error threshold (ASR)

Source: compiled by the authors

The diagram demonstrates that the ASR module is the bottleneck in the system; any reduction in WER will enhance the reliability of the remaining components of the system. Achieving a WER of between 5 and 7% with models such as Whisper or other cloud-based offerings creates new conditions for building reliable systems for the detection of informational influences. Thus, the analysis of each of the models presented led to the conclusion that neural network models have made a radical improvement in the accuracy of ASR systems compared to classical models. The enhanced accuracy of end-to-end neural network ASR models has led to an improvement in the quality of transcription and the reliability of the remaining components of the automated analysis system. Consequently, such systems can be deployed in practice to automate the analysis of informational content.

Analysis of the computational performance and time characteristics of ASR systems

One of the key parameters in assessing the suitability of a system for practical applications is its computational performance. The computational performance of an ASR system can be reflected in the processing time of the audio signals that are processed by the system. A comparison of the CPU-oriented classical systems (HTK, CMU Sphinx, Kaldi) and the GPU-optimised DL models (DeepSpeech, Wav2Vec, Whisper) reveals the relationship between the hardware that supports an ASR system and its processing time for an audio file. Classic systems, which traditionally run on a CPU, demonstrate the longest processing times. For example, HTK required 72 seconds to transcribe a standard benchmark file, CMU Sphinx – 51 seconds, and Kaldi – 61 seconds. These figures are explained by the sequential architecture of HMM-based models, where the computation of acoustic and linguistic probabilities takes place on the CPU, without significant use of the parallel computations' characteristic of GPU. In these cases, processing time is determined not only by the complexity of the algorithms, but also by the computational throughput limitations of the central core, which significantly slows down processing of long or multi-channel audio streams.

Modern DL-based models, by contrast, demonstrate significant speed-ups thanks to optimisations for GPU computing. DeepSpeech processes a benchmark file in 38 seconds, Wav2Vec in 29 seconds, and Whisper Large-v3 in just 10 seconds, effectively delivering processing speeds faster than real time. This difference is attributed both to the parallel processing of tensor operations on the GPU and to the integration of acoustic and language modelling stages into a single end-to-end architecture, which eliminates the cascading accumulation of errors and additional intermediate computations characteristic of classical systems. Thanks to the use of a transformer architecture with multi-level self-learning blocks for speech representations, this model achieves a low WER (7%) while requiring minimal processing time. However, the use of smaller models within the Whisper library results in an increase in the WER to 10-15%. Thus, local speech recognition models with a GPU seem to be the most effective solution for tasks that require high-speed speech recognition. CPU-based solutions are still relevant for applications that require the use of stable algorithms that do not make use of GPU, or for the processing of large volumes of data that can be performed offline.

A comparison of local and cloud-based ASR systems

A comparison of local speech recognition systems to cloud-based speech recognition APIs reveals some

important differences between the two solutions. Cloud API providers, such as Google and Microsoft, offer speech recognition services that have comparable WERs to the best of the local speech recognition models. However, a closer look at specific metrics for both local and cloud-based systems reveals the true difference between the two methods of speech recognition. Results of the comparison between the local speech recognition models and the cloud APIs are presented in Table 2.

Table 2. A comparison of on-premises models and cloud services in terms of WER and processing time

Category	System	WER, %	Processing time
Local DL	Whisper Large-v3	7	~10 s
	Wav2Vec	12	~29 s
	DeepSpeech	22	~38 s
Cloud API	Google STT	7	~10-20 s
	Azure Speech	7	~5-10 s
	Amazon Transcribe	10	~10-15 s
	IBM Watson STT	15	~10-15 s

Source: compiled by the authors

Results presented show that cloud solutions can be compared with on-premises solutions in terms of the achieved performance metrics. However, there are differences between the two solutions in relation to the latency of the system. For local systems, the latency is only associated with the computational power of the system's hardware. For cloud solutions, end-to-end latency is associated with the transmission of data to third-party servers, which can introduce a stochastic component into the system's latency. In comparison with a local system whose latency is determined by the system's hardware, end-to-end latency for a cloud system also depends on the performance of the provider's servers and network. During periods of high demand for the provider's services or during network issues, end-to-step latency for cloud services can significantly increase, making those solutions unsuitable for specific applications that require high levels of real-time processing and analysis of the data, such as monitoring information flows or audio content in real time.

Cloud solutions require the transmission of audio files to third-party servers to be stored or used for training models as prescribed by the cloud service provider. This introduces additional risks for companies whose audio files contain sensitive information, such as business or personal communications, meetings, or journalistic interviews. The use of local solutions allows the company to maintain full control over its data (Jin *et al.*, 2025). For applications requiring high levels of real-time processing, on-premises solutions have a significant advantage over cloud solutions. The average latency for both solutions is similar, but the jitter (variation in latency) for cloud solutions is significantly higher. Local systems use on-premises models that are optimised for GPUs, which allows them to produce more stable results than cloud solutions (Yang *et al.*, 2024). Therefore, while the accuracy of cloud APIs and on-premises models is comparable, cloud solutions are significantly less suitable for applications that require a high level of availability, low latency, and data control. Solutions based on the Whisper and Wav2Vec models for on-premises systems, which offer a good balance of accuracy, performance, and security, present themselves as a promising solution for building autonomous systems for the analysis of audio data containing sensitive information.

Noise robustness and linguistic characteristics of Ukrainian data

An analysis of speech recognition systems in the context of the Ukrainian language and its use in a bilingual environment has demonstrated how the features of the language have a significant influence upon the accuracy of the recognition systems. The Ukrainian language contains soft consonants, which significantly increases the difficulty of recognition systems whose acoustic models do not account for these sounds. These sounds and other language features significantly complicate the process of segmenting phonemes in recognition systems whose models are based upon HMMs and other statistical features of speech. Furthermore, the language's use in a bilingual environment, especially with Russian, introduces further challenges to recognition systems based upon these classical models. These challenges result from the use of code-switching (switching between languages within the same sentence), loanwords, and accents from different regions of Russia. Additionally, the classical models have a limited training corpus that does not contain enough examples of bilingual speech to effectively recognise the sounds and linguistic features associated with those languages when spoken in conjunction. Therefore, to increase the accuracy of classical models to Ukrainian speech, it would be necessary to either expand the training corpus to include examples of speech with these features, or to perform fine-tuning of the models to account for these specific linguistic features (Bai & Katkov, 2025).

Another reason why classical acoustic models based upon HMMs have low accuracy in the recognition of the Ukrainian language is due to the shortage of training corpora that contain speech examples in that language. Furthermore, these classical models focus upon analysing statistical features of speech within limited ranges

of the speech signal, which prevents those models from generalising well to other types of speech signals. As a result of these factors, classical systems reveal errors (WER) in recognising the Ukrainian language between 25% and 30%, which is a significant error rate for information flow and keyword detection analysis systems. In contrast, neural network models such as Wav2Vec and Whisper exhibit significantly greater error tolerance for language features and noisy environments. These models have been trained on a large number of speech signal examples that include bilingual speech. The ability of these models to learn the representations of phonemes in the Ukrainian language and to understand the language's stress and soft consonants allows the models to recognise those sounds correctly. Furthermore, the exposure of these models to bilingual speech during training allows the models to recognise code-switching between Ukrainian and Russian speech, and to reduce errors that result from that code-switching. Experiments with the Wav2Vec and Whisper models reveal WER rates for Ukrainian-Russian speech signals between 7% and 12%, which is between 2 and 3 times lower than that of classical recognition systems. The results of these experiments are presented in Table 3.

Table 3. Impact of linguistic factors on WER for different classes of systems

Factor	Classical systems (HTK, CMU Sphinx)	Local DL (Wav2Vec, Whisper)	Cloud API
Soft consonants	28-30%	7-12%	7-10%
Stress variations	25-28%	7-12%	7-10%
Bilingual environment	30-32%	10-12%	7-10%
Background noise	27-31%	8-12%	8-10%

Source: compiled by the authors

Thus, the results of this study showed that in order to effectively recognise the Ukrainian language and speech of bilingual individuals, the use of language models that were pre-trained on extensive corpora of diverse and Ukrainian language texts is required. The presence of multilingual and noisy data during the training of these language models ensures that the models have developed a stable representation of speech sounds, allowing them to account for variations in the stress of syllables, the use of soft consonants, and code-switching between languages. Furthermore, the use of multilingual neural network models, especially end-to-end transformer models for ASR, has shown a low WER even in the presence of noise in the recorded audio files.

Assessment of hardware requirements and computational complexity of ASR systems

Another of the main priorities in the selection of ASR systems for various applications is the assessment of the hardware requirements of the ASR systems. The hardware requirements of a computing system can have a major impact upon the processing time of the system's audio files, and upon whether the system is able to be deployed in the targeted application of the ASR system. For instance, ASR systems based upon frameworks like HTK, CMU Sphinx, and Kaldi have low hardware requirements for their operation; they are based upon algorithms that are performed by the computer's CPU rather than relying upon the computational power of the computer's graphics card.

However, this determinism comes at the cost of significant processing time, particularly when processing long audio streams or multi-channel signals, which can be critical for real-time applications. Modern DL systems, including DeepSpeech, Wav2Vec and Whisper, exhibit a fundamentally different hardware utilisation profile. Thanks to transformer and convolutional architectures, these models make intensive use of GPUs for parallel processing of tensor operations, enabling processing times an order of magnitude lower than in classical systems, with significantly higher accuracy.

The practical implications for deploying systems on edge devices or in local on-premise environments are reflected in the choice of model, depending on the hardware platform's limitations. Local edge devices with limited GPU power effectively integrate smaller-scale models (Tiny or Base), which, although they have a slightly higher WER (10-15%), provide real-time processing and reduced power consumption. On enterprise servers or on-premise deployments with high-power graphics cards, it is advisable to use Large models, such as Whisper Large-v3 or Wav2Vec, which combine low WER with high processing speed – a critical factor for real-time analytics systems and complex streaming tasks. Options for optimising hardware requirements include a trade-off between speed and accuracy, model selection, and the integration of compression and acceleration techniques such as pruning, distillation, or quantised representations (Table 4). Such approaches make it possible to adapt large neural network models to limited resources without a significant loss of accuracy, making them suitable for both edge devices and server-based solutions where minimising response times and conserving resources are key.

A comprehensive analysis of the results of speech recognition system evaluations indicates that effective detection of the information effect is only possible through a holistic approach that simultaneously covers accuracy, performance and the specific characteristics of the input data. Classic CPU-oriented models, such as HTK, CMU Sphinx and Kaldi, whilst providing deterministic processing, demonstrate limited suitability for analytical scenarios with highly dynamic audio streams,

as their accuracy (25-30% WER) and processing time (51-72 seconds per standard file) significantly limit the capabilities of automated classification, sentiment analysis and keyword search. Such characteristics make them primarily relevant for offline analysis of archived data or research purposes.

Table 4. Hardware requirements and performance for different system classes

System/Model	CPU	GPU	Processing time (s)	WER (%)	Suitability for edge/on-premise
HTK	+	-	72	25	limited (offline)
CMU Sphinx	+	-	51	30	limited (offline)
Kaldi	+	-	61	18	limited (offline)
DeepSpeech	+	+	38	22	on-premise, highly efficient
Wav2Vec	+	+	29	12	Edge/server use
Whisper Tiny	+	+	5-8	10	edge-devices
Whisper Base	+	+	7-10	8-10	edge/on-premise
Whisper Large-v3	+	+	10	7	on-premise, real-time analytics

Source: compiled by the authors

Local DL models, notably Wav2Vec and Whisper, demonstrate an optimal balance of performance and accuracy. Wav2Vec achieves a WER of around 12% whilst taking 29 seconds to process the audio file. Whisper Large-v3 has a minimum WER of 7% but takes only 10 seconds to process the same file, making it more suitable for applications requiring faster response times, the analysis of larger volumes of audio data, and which aim to minimise errors within the analytical process. Each of these platforms has been selected for its abilities to recognise multiple languages, phonetic variations within the Ukrainian language, and for its ability to process noisy audio files. Furthermore, the models that are optimised for GPUs can be deployed both at the edge and on-premise. Finally, whilst commercial cloud services like Google STT, Azure Speech and Amazon Transcribe also report WER values of between 7 and 10%, their latency to respond to requests, their potential for performance errors during peak times, and their data policies regarding the processing of personal data can present obstacles to their implementation within systems requiring rapid responses from the service and which require control over the data being analysed. A generalised comparison of system classes based on key parameters is presented in Table 5.

Table 5. A general comparison of system classes based on key parameters

System class	Accuracy (WER, %)	Processing time	Hardware requirements	Noise suppression capabilities	Suitability for analytical tasks
Traditional CPU	25-30	51-72 s	CPU	Average	limited, offline
Local DL (Wav2Vec)	12	29 s	CPU+GPU	High	edge/on-premise, real-time
Local DL (Whisper Large-v3)	7	10 s	CPU+GPU	Very high	edge/on-premise, real-time analytics
Cloud API	7-10	~5-20 s (depending on the network)	-	High	analytics, but infrastructure-dependent

Source: compiled by the authors

Based on the analysis, a range of practical conclusions can be formed. For tasks involving information extraction and automated analysis of text derived from audio, it is advisable to base the system on local neural network models, such as Whisper and Wav2Vec, which ensure error minimisation, noise robustness and the ability to integrate into confidential environments. The choice of a specific model must balance model size and performance: large models deliver the lowest WER but require significant computational resources, whilst smaller models can be used for rapid data processing on edge devices at the cost of a slight increase in WER. Key constraints for system implementation include hardware infrastructure, the availability of language and acoustic corpora, and the scalability of processing for large streams of audio data. Thus, the integration of local multilingual models into analytical systems provides an optimal balance between performance, accuracy and data control, creating prospects for further research in the field of automated detection of information effects and systems for real-time monitoring of speech streams.

DISCUSSION

An analysis of academic publications from 2022 to 2025 has revealed active development in ASR methods, particularly for multilingual and low-resource languages. The paper by D.A. Zimenkov & I.A. Tereikovskyi (2025) investigates the challenges of multilingual ASR systems, specifically the limited amount of annotated data,

phonetic variability in speech, the impact of acoustic noise, and differences in the intonational and rhythmic characteristics of languages. The authors applied transformer architectures and transfer learning methods, which reduce WER and increase the models' robustness to acoustic distortions and noise. Compared to the study in question, which tested local and cloud-based systems on a Ukrainian corpus with the addition of white noise at an SNR of 10 dB, the authors' approach demonstrated a similarly enhanced resistance to noise. At the same time, J.L.K.E. Fendji *et al.* (2022) showed that integrating acoustic and linguistic modelling into a single architecture simplified the signal processing pipeline and improved recognition accuracy. The study also observed that end-to-end models (DeepSpeech, Wav2Vec 2.0, Whisper) outperformed classical HMM-GMM systems in terms of accuracy and noise robustness; however, the authors' results highlighted the importance of customising dictionaries for specialised terminology. In turn, current experiments with the Ukrainian language have shown that pre-trained models may require additional fine-tuning for locally specific corpora.

A comprehensive review of models for regional languages was conducted by S. Shaikh *et al.* (2024). The authors' experiments demonstrated that self-supervised learning and fine-tuning substantially improved accuracy for low-resource languages. A similar trend was observed in the current study: Wav2Vec 2.0, which utilises self-supervised learning on large corpora, demonstrated improved accuracy in Ukrainian language recognition compared to other models. A. Senyk (2025) highlighted the importance of fine-tuning models for high-quality language recognition and generation. In the present study, local modifications and fine-tuning of models on Ukrainian data improved accuracy and reduced WER compared to "clean" pre-trained models. Meanwhile, C. Chen's (2025) thesis conducted an in-depth study of strategies for adapting large ASR models to specific linguistic and acoustic conditions, including fine-tuning and transfer learning methods. The author demonstrated that adaptation significantly reduced WER in low-resource scenarios and improved the models' robustness to noise. Current experiments have shown that additional local fine-tuning on Ukrainian corpora yielded noticeably better accuracy for short recordings, which can be explained by linguistic features and the size of the training dataset.

The accuracy and adaptability of end-to-end ASR models in complex linguistic contexts were evaluated by P. Jain & A. Bhowmick (2025). They demonstrated that self-supervised learning on large unlabelled corpora, followed by fine-tuning, significantly improved model performance, particularly in languages with limited resources. Similar results were also obtained in the study presented in this paper. Wav2Vec 2.0, which was trained on multilingual data and later trained on Ukrainian recordings, also exhibited a significant reduction in WER compared to the baseline models, confirming the universality of these self-supervised learning approaches. Another Turkish language recognition system was also implemented by H. Polat *et al.* (2024). The authors found that the use of Low-Rank Adaptation (LoRA) allowed their system to significantly reduce the computational costs needed to train the model while maintaining accuracy, and that fine-tuning the model on Turkish audio corpora improved the model's noise robustness. Current results correlate with this conclusion: the application of Whisper with fine-tuning on Ukrainian data improved accuracy and robustness to noise distortions. The significance of standardised data processing pipelines and multi-criteria evaluation schemes for comparing models across different linguistic and acoustic contexts was highlighted by M. Trenton (2025) in a comprehensive review of multilingual and multimodal ASR systems. The study adopted a similar approach, creating a unified audio processing pipeline that incorporates volume normalisation, VAD segmentation, and a multi-criteria evaluation of accuracy, noise robustness and computational requirements.

M. Si *et al.* (2024) proposed the Lexical Error Guard method, demonstrating that integrating an LLM into post-processing reduced the number of insertions, deletions and word substitutions compared to baseline algorithms, thereby improving the overall ASR accuracy. Compared to the presented study, which employed Wav2Vec 2.0 and Whisper for the Ukrainian language and evaluated accuracy on test audio recordings with white noise added at a SNR of 10 dB, the results of Lexical Error Guard demonstrated a similar reduction in WER; however, this approach addressed local fine-tuning of models without additional integration of large language models for post-processing, which was reflected in slightly higher error rates for complex words and numerical entities. A. Bhandari & G. Harit (2026) demonstrated that correction algorithms significantly improved accuracy in short phrases and reduced errors associated with phonetically similar words. A similar effect was observed in the presented study: additional text normalisation and punctuation restoration rules reduced the WER for short entries, confirming the universality of post-processing approaches for low-resource languages.

Wav2Vec 2.0 was the subject of research by G. Shekoufandeh (2023) and Q. Li (2024). G. Shekoufandeh found that age and speech characteristics had a significant impact on model accuracy, whilst additional fine-tuning on specific corpora substantially improved the results. Meanwhile, Q. Li supplemented these findings and demonstrated that retraining on low-resource language corpora, considering phonetic differences between related languages, can reduce WER and improve word segmentation accuracy. Compared to the study in question, which utilised fine-tuning of Wav2Vec 2.0 and Whisper on Ukrainian corpora, the authors' results

confirmed the effectiveness of the fine-tuning approach for low-resource languages, but also demonstrated that success depended on the degree of phonetic similarity between the source and target languages.

V. Delić *et al*. (2019) conducted a retrospective review of the development of TTS and STT technologies for South Slavic languages with limited resources. They emphasised the importance of multilingual and multimodal architectures, standardised data processing pipelines, and multi-criteria performance evaluation schemes for improving the accuracy and robustness of models. S. Nithya *et al*. (2025) applied multimodal approaches and specialised data normalisation strategies to NLP, which significantly improved performance in noisy environments. Compared to the present study, which evaluated local and cloud-based STT models for the Ukrainian language, the study determined that such multimodal approaches provide increased robustness, but they require significantly greater computational resources and a more complex data processing pipeline, whereas the current experiments addressed the optimisation of local models with controlled resources. At the same time, Y. Li *et al*. (2025) noted that for low-resource dialects, the availability of well-structured training corpora and effective text post-processing to correct typical errors is critical.

A study by S. Saranya *et al*. (2025) showed that combining ASR with an automatic abstract summarisation system reduced the error rate whilst increasing the practical benefit for the end user. The authors' results demonstrated expanded practical applications of STT, but emphasised that additional modules can introduce delays and require extra computational resources, which explains the partial lack of correlation in processing speed between studies. These results were expanded upon by K. Fatehi (2023), who showed that models pre-trained on large unlabelled audio corpora and then fine-tuned on limited labelled data demonstrated a significant reduction in WER and increased robustness to noise. Compared to the presented study, the author's approach confirmed the effectiveness of self-supervised learning for low-resource languages; however, it required significantly greater computational resources and more complex audio pre-processing strategies.

S.M.H. Amiri (2025) investigated multilingual NLP systems and speech recognition for global communication, integrating various ASR models into multi-domain scenarios. The author demonstrated that combining different architectures and using language-specific adaptation improved accuracy for languages with limited data, whilst reducing the need for large annotated corpora. In turn, S.M. Zahid & S.A. Qazi (2025) showed that variations in pitch and speech rate during training significantly improve the models' ability to generalise to new recordings and reduce the number of common errors. Compared to the presented study, the results of the studies reviewed indicated a potential path to improving accuracy through generative data augmentation methods, which explains why current models demonstrated a higher WER in some scenarios. Y. Yang *et al*. (2025) presented GigaSpeech 2 – a large multilingual corpus for low-resource languages with automated collection, transcription and post-processing of audio data. The study demonstrated that large and diverse corpora improved performance of end-to-end ASR models, particularly in noisy and multilingual environments. Compared to the study in question, the results by Y. Yang *et al*. confirmed the importance of data scale and diversity, which explains the difference in the scalability of systems and their performance on large and heterogeneous corpora. Thus, the analysis of the literature confirmed the key role of self-supervised learning, multilingual adaptation, generative data augmentation and large corpora in improving the accuracy and robustness of ASR in resource-constrained settings. A comparison with the current study showed that the local and cloud-based Wav2Vec 2.0 and Whisper models delivered stable performance for the Ukrainian language, but differences in data scale, the use of self-supervised learning and multilingual architectures accounted for some discrepancies in accuracy and processing speed metrics.

CONCLUSIONS

The study found that ASR performance depended significantly on the model architecture and the computational pipeline. Classic statistical systems based on HMM (HTK, CMU Sphinx, Kaldi) demonstrated the worst performance, with WERs ranging from 18-30% and processing times of 51-72 seconds for the reference audio file. End-to-end neural network models achieved significantly higher accuracy: DeepSpeech achieved a WER of ≈22%, Wav2Vec – ≈12%, and Whisper Large-v3 – ≈7%. Thus, the transition from traditional cascade approaches to modern transformer architectures has reduced the number of errors by a factor of 3-4. It was also confirmed that even a slight increase in WER leads to a noticeable deterioration in the results of subsequent NLP tasks (classification, sentiment analysis, keyword search), which highlights the critical role of the ASR module in the entire analytical pipeline. Experimental evaluation under noise conditions and with Ukrainian code-switching showed that classical acoustic models achieved a WER of 25-32%, whilst multilingual transformer architectures achieved 7-12%. Performance analysis revealed a significant advantage for GPU-optimised local models: Whisper Large-v3 processed the reference audio file in approximately 10 seconds, Wav2Vec in 29 seconds, whilst classical systems required over 60 seconds. Cloud-based ASR services demonstrated comparable accuracy (7-10% WER) but were characterised by variable response times due to network delays and load.

In summary, the study's findings confirmed that modern local end-to-end neural network models, particularly Whisper and Wav2Vec, strike an optimal balance between accuracy, speed and data control. Achieving

a WER of 5-7% created qualitatively new conditions for building reliable systems for analysing audio content and detecting information effects, as the improvement in ASR accuracy directly enhanced all subsequent processing stages. The study demonstrated that the transition to deep end-to-end architectures is a key prerequisite for the practical use of such systems in real-world multilingual and noisy information environments. The main limitation of this study is the dependence of the results on the availability of representative Ukrainian and bilingual audio corpora, as well as on hardware resources, which limit the scale of the experiments and the comprehensiveness of model evaluation in various real-world scenarios. Further research should expand specialised Ukrainian and code-switching corpora, integrating adaptive fine-tuning and edge-level optimisations, as well as conducting an in-depth analysis of the impact of ASR errors on the final metrics of NLP components in information effect detection systems.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Amazon Transcribe. (n.d.). Retrieved from https://aws.amazon.com/transcribe.

[2] Amiri, S.M.H. (2025). Beyond language barriers: Multilingual NLP and voice recognition for global connectivity. *International Journal of Science and Research Archive*, 15(2), 406-419. doi: 10.30574/ijsra.2025.15.2.1346.

[3] Azure Speech in Foundry Tools. (n.d.). Retrieved from https://azure.microsoft.com/en-us/products/ai-services/ai-speech.

[4] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th international conference on neural information processing systems* (pp. 12449-12460). Red Hook: Curran Associates Inc. doi: 10.5555/3495724.3496768.

[5] Bai, Y.V., & Katkov, Yu.I. (2025). Review of research on non-audio recognition using neural networks. *Scientific Notes of the State University of Information and Communication Technologies*, 1(7), 106-115. doi: 10.31673/2786-8362.2025.018603.

[6] Bhandari, A., & Harit, G. (2026). Post-ASR correction for low-resource Rajasthani language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 25(3), article number 15. doi: 10.1145/3793254.

[7] Bouchelligua, W. (2025). VoiceX: An end-to-end speech recognition model for low-resource languages using transfer learning. *Arabian Journal for Science and Engineering*, 51, 11893-11920. doi: 10.1007/s13369-025-10859-7.

[8] Chen, C. (2025). *Advancing speech-to-text adaptation for large speech models*. Singapore: Nanyang Technological University. doi: 10.32657/10356/201854.

[9] CMU Sphinx. (n.d.). Retrieved from https://github.com/cmusphinx.

[10] DeepSpeech. (n.d.). Retrieved from https://github.com/mozilla/DeepSpeech.

[11] Delić, V., Perić, Z., Sečujski, M., Jakovljević, N., Nikolić, J., Mišković, D., Simić, N., Suzić, S., & Delić, T. (2019). Speech technology progress based on new machine learning paradigm. *Computational Intelligence and Neuroscience*, 2019(1), article number 4368036. doi: 10.1155/2019/4368036.

[12] Dhahbi, S., Saleem, N., Bourouis, S., Berrima, M., & Verdú, E. (2025). End-to-end neural automatic speech recognition system for low resource languages. *Egyptian Informatics Journal*, 29, article number 100615. doi: 10.1016/j.eij.2025.100615.

[13] Dumyn, A., Basystiuk, O., Dumyn, I., & Shakhovska, N. (2025). Performance evaluation of automatic speech recognition systems for Ukrainian and English languages. In P. Štarchoň, S. Fedushko & K. Gubíniova (Eds.), *Developments in information and knowledge management systems for business applications* (pp. 423-443). Cham: Springer. doi: 10.1007/978-3-031-95955-4_23.

[14] Fatehi, K. (2023). *Self-supervised learning for automatic speech recognition in low-resource environments*. Nottingham: University of Nottingham.

[15] Fendji, J.L.K.E., Tala, D.C.M., Yenke, B.O., & Atemkeng, M. (2022). Automatic speech recognition using limited vocabulary: A survey. *Applied Artificial Intelligence*, 36(1), article number 2095039. doi: 10.1080/08839514.2022.2095039.

[16] Froiz-Miguez, I., Fraga-Lamas, P., & Fernández-CaraméS, T.M. (2023). Design, implementation, and practical evaluation of a voice recognition based IoT home automation system for low-resource languages and resource-constrained edge IoT devices: A system for Galician and mobile opportunistic scenarios. *IEEE Access*, 11, 63623-63649. doi: 10.1109/ACCESS.2023.3286391.

[17] Google Cloud. (n.d.). *Cloud Speech-to-Text overview*. Retrieved from https://cloud.google.com/speech-to-text/docs/speech-to-text-requests.

[18] Hidden Markov Model Toolkit. (n.d.). Retrieved from https://github.com/mxochicale/htk.

[19] IBM Watson Speech to Text. (n.d.). Retrieved from https://www.ibm.com/products/speech-to-text.

[20] Jain, P., & Bhowmick, A. (2025). Comparative performance analysis of end-to-end ASR models on Indo-Aryan and Dravidian languages within India's linguistic landscape. *EURASIP Journal on Audio, Speech, and Music Processing*, 2025(1), article number 10. doi: 10.1186/s13636-025-00395-5.

[21] Jin, W., Cao, Y., Su, J., Wang, D., Zhang, Y., Xue, M., Hao, J., Dong, J.S., & Yang, Y. (2025). Whispering under the eaves: Protecting user privacy against commercial and LLM-powered automatic speech recognition systems. *ArXiv*. doi: 10.48550/arXiv.2504.00858.

[22] Kaldi. (n.d.). Retrieved from https://github.com/kaldi-asr/kaldi.
[23] Kumar, L.A., Renuka, D.K., Chakravarthi, B.R., & Mandl, T. (2024). *Automatic speech recognition and translation for low resource languages*. Hoboken: John Wiley & Sons. doi: 10.1002/9781394214624.
[24] Li, Q. (2024). *Fine-tuning Cantonese based on Wav2vec 2.0 XLRS model that pretrained on Mandarin Chinese to improve ASR performance*. Groningen: University of Groningen.
[25] Li, Y., Wang, Y., Hoi, L.M., Yang, D., & Im, S.-K. (2025). A review on speech recognition approaches and challenges for Portuguese: Exploring the feasibility of fine-tuning large-scale end-to-end models. *EURASIP Journal on Audio, Speech, and Music Processing*, 2025(1), article number 3. doi: 10.1186/s13636-024-00388-w.
[26] Liu, W. (2025). *Enhancing efficiency and reliability in automatic speech recognition systems*. Shatin: The Chinese University of Hong Kong.
[27] Mujtaba, D.F., & Mahapatra, N.R. (2025). Automatic speech recognition APIs: Models, features, and challenges. In H.R. Arabnia, L. Deligiannidis, F. Shenavarmasouleh, S. Amirian & F.G. Mohammadi (Eds.), *11th international conference: Computational science and computational intelligence* (pp. 33-48). Cham: Springer. doi: 10.1007/978-3-031-94940-1_3.
[28] Nithya, S., Sengar, A.S., Jayalakshmi, K., Kokulavani, K., Joel, P.J., & Srinivasulu, M. (2025). Multilingual and robust speech recognition: Leveraging advanced machine learning for accurate and realtime natural language processing. In *Proceedings of the 1st international conference on research and development in information, communication, and computing technologies* (pp. 513-518). Nagapattinam: SciTePress. doi: 10.5220/0013885700004919.
[29] OpenAI. (2022). *Whisper*. Retrieved from https://openai.com/uk-UA/index/whisper.
[30] Polat, H., Turan, A.K., Koçak, C., & Ulaş, H.B. (2024). Implementation of a whisper architecture-based Turkish automatic speech recognition (ASR) system and evaluation of the effect of fine-tuning with a low-rank adaptation (LoRA) adapter on its performance. *Electronics*, 13(21), article number 4227. doi: 10.3390/electronics13214227.
[31] Rakhimova, D., Duisenbekkyzy, Z., & Adali, E. (2025). Investigation of ASR models for low-resource Kazakh child speech: Corpus development, model adaptation, and evaluation. *Applied Sciences*, 15(16), article number 8989. doi: 10.3390/app15168989.
[32] Saranya, S., Bharathi, B., Dhanya, S.G., & Krishnakumar, A. (2025). Real-time continuous Tamil dialect speech recognition and summarization. *Circuits, Systems, and Signal Processing*, 44(4), 2855-2881. doi: 10.1007/s00034-024-02950-5.
[33] Senyk, A. (2025). *Implementing text-to-speech synthesis with voice cloning for the Ukrainian language*. Lviv: Ukrainian Catholic University.
[34] Shaikh, S., Pereira, K.W., Sahay, S., Lopes, A., & Parshionikar, S. (2024). An extensive review: Models for regional language speech recognition. In *2024 4th Asian conference on innovation in technology* (pp. 1-8). Pimari Chinchwad: IEEE. doi: 10.1109/ASIANCON62057.2024.10837903.
[35] Shekoufandeh, G. (2023) *Evaluation of wav2vec 2.0 speech recognition for the elderly Frisian population*. Groningen: University of Groningen.
[36] Si, M., Cobas, O., & Fababeir, M. (2024). Lexical error guard: Leveraging large language models for enhanced ASR error correction. *Machine Learning and Knowledge Extraction*, 6(4), 2435-2446. doi: 10.3390/make6040120.
[37] Trenton, M. (2025). A comprehensive survey on multilingual and multimodal automatic speech recognition systems. *Journal of Computer Technology and Software*, 4(10).
[38] Yang, B., Hu, Q., Xie, W., Wang, X., Luo, W., & Zhang, Q. (2024). PDAssess: A privacy-preserving free-speech based Parkinson's disease daily assessment system. In *Proceedings of the 21st ACM conference on embedded networked sensor systems* (pp. 251-264). New York: Association for Computing Machinery. doi: 10.1145/3625687.3625805.
[39] Yang, Y., *et al*. (2025). GigaSpeech 2: An evolving, large-scale and multi-domain ASR corpus for low-resource languages with automated crawling, transcription and refinement. In *Proceedings of the 63rd annual meeting of the association for computational linguistics* (Vol. 1; pp. 2673-2686). Vienna: Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.135.
[40] Zahid, S.M., & Qazi, S.A. (2025). Pitch-speed feature space data augmentation for automatic speech recognition improvement in low-resource scenario. *IEEE Access,* 13, 86998-87014. doi: 10.1109/ACCESS.2025.3569579.
[41] Zimenkov, D.A., & Tereikovskyi, I.A. (2025). Analysis of problems in speech recognition methods. *Scientific Notes of the V.I. Vernadsky TNU. Series: Technical Sciences*, 36(75(3)), 198-205. doi: 10.32782/2663-5941/2025.3.2/27.

Аналіз методів та підходів до задачі обробки аудіо даних

Владислав Радін

Аспірант
Державний університет "Київський авіаційний інститут"
03058, просп. Любомира Гузара, 1, м. Київ, Україна
https://orcid.org/0009-0009-1101-6888

Мирослав Рябий

Кандидат технічних наук, доцент
Державний університет "Київський авіаційний інститут"
03058, просп. Любомира Гузара, 1, м. Київ, Україна
https://orcid.org/0000-0002-9651-9135

Анотація. Актуальність дослідження визначається потребою комплексної оцінки сучасних систем автоматичного розпізнавання мовлення в умовах української мовної специфіки з урахуванням компромісу між точністю, продуктивністю, апаратними вимогами та впливом якості транскрипції на подальші інформаційно-аналітичні процеси. Мета дослідження полягала в порівнянні наявних методів і технологій розпізнавання мовлення для побудови системи оцінки інформаційного впливу. У дослідженні використовувалися традиційні моделі прихованих марковських процесів та гауссові суміші, гібридні нейронні мережі, локальні нейромережеві моделі DeepSpeech, Wav2Vec 2.0 і Whisper, хмарні сервіси розпізнавання мовлення від Google, Amazon, Microsoft та IBM. Здійснювалася обробка аудіо з використанням виявлення мовлення та мел-частотних коефіцієнтів, фіксувалися точність розпізнавання слів, час обробки на процесорі та графічному процесорі, затримка та нестабільність відповіді, а також оцінювався вплив помилок розпізнавання на конвеєр обробки природної мови з токенізацією, тематичною класифікацією та аналізом настрою. У ході дослідження встановлено істотну залежність точності та продуктивності Automatic Speech Recognition від архітектури моделей і апаратної підтримки. Класичні Hidden Markov Models продемонстрували найнижчу ефективність із Word Error Rate 18-30 % та часом обробки 51-72 с, що підтвердило їх обмежену придатність для української мови в умовах шуму й двомовного мовлення. Локальні end-to-end нейромережеві моделі забезпечили принципово кращі показники: DeepSpeech досяг 22 % Word Error Rate, Wav2Vec – близько 12 %, а Whisper Large-v3 – приблизно 7 %, що відповідає зменшенню кількості помилок у 3-4 рази. Було показано, що багатомовні трансформерні архітектури характеризуються підвищеною стійкістю до фонетичних варіацій української мови та код-свічингу. Оптимізовані локальні моделі забезпечили обробку майже в реальному часі (Whisper Large-v3 ≈10 с на файл), суттєво перевершивши класичні підходи за швидкодією. Хмарні сервіси продемонстрували співставну точність (7-10 % Word Error Rate), однак виявили залежність від мережевих затримок і обмеження щодо конфіденційності. Узагальнено, Whisper і Wav2Vec визначено як оптимальну основу для систем аналізу аудіоконтенту та виявлення інформаційного ефекту, оскільки вони забезпечують найкращий баланс між точністю, продуктивністю та контролем над даними. Отримані результати можуть бути застосовані у розробці систем автоматичної транскрипції, моніторингу та аналізу аудіопотоків, розробці сервісів голосової аналітики та модулів раннього виявлення інформаційних впливів

Ключові слова: speech-to-text; автоматичне розпізнавання мовлення; Whisper; Wav2Vec; обробка аудіопотоків; Word Error Rate; нейронні мережі

BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 2. 2026

UDC 004.253
DOI: 10.62660/bcstu/2.2026.48

ISSN: 2306-4412
E-ISSN: 2708-6070

Rethinking memory hierarchy policies in DRAM+HBM systems

Maksym Shcherba*
Postgraduate Student
Ivan Franko National University of Lviv
79007, 1 Universytetska Str., Lviv, Ukraine
https://orcid.org/0009-0008-0332-9254

Petro Tarnavskyi
Postgraduate Student
Ivan Franko National University of Lviv
79007, 1 Universytetska Str., Lviv, Ukraine
https://orcid.org/0000-0003-3265-8168

Abstract. Efficient management of heterogeneous memory systems integrating high-bandwidth memory (HBM) is critical for overall performance. Conventional approaches often assume that dynamic random-access memory (DRAM) is the fastest tier, which may not hold true in HBM-equipped configurations. This study aimed to evaluate the performance impact of reconfiguring memory hierarchy policies by shifting the preference from DRAM to HBM in DRAM+HBM systems. The research employed a modified Ambix framework, adapting migration policies to designate HBM as the fast tier and tuning internal parameters to better suit the DRAM+HBM architecture. Performance was assessed using HPCG and AMG benchmarks on a dual-socket server. It was established that placing all data in HBM yielded performance gains of up to 24% over DRAM-only execution. The analysis demonstrated that standard Linux NUMA balancing could cause up to 4% performance degradation because it incorrectly promoted memory pages to the slower DRAM tier. Investigations revealed that while default Ambix configurations might reduce performance by up to 23%, favouring HBM improved results by 1.4% to 22% over DRAM-preferred policies. Furthermore, internal parameters were tuned to reduce page table scanning frequency and increase migration limits per cycle. These modifications achieved a 3-7% performance gain for the HPCG benchmark compared to the baseline system. For the AMG benchmark, experimental results showed variations from 1.5% degradation to 1.4% improvement, depending on the DRAM:HBM capacity ratios. These findings can enable system architects to optimise memory tiering in next-generation servers without requiring kernel or application modification

Keywords: heterogeneous memory systems; memory management; dynamic random-access memory; high bandwidth memory; memory hierarchy

INTRODUCTION

The ongoing evolution of high-performance computing is increasingly characterised by the integration of diverse memory technologies, such as high-bandwidth memory (HBM), alongside conventional dynamic random-access memory (DRAM). This development offers substantial potential for performance gains, particularly for data-intensive applications that require high throughput. However, a notable consideration in these environments is that many existing system management policies were originally designed for traditional architectures where DRAM was the undisputed fast tier. When these established mechanisms are applied to mixed DRAM and HBM configurations, they may not fully exploit the specific performance characteristics

Article's History: Received: 08.01.2025; Revised: 17.04.2026; Accepted: 18.05.2026; Published: 26.06.2026

Suggested Citation:
Shcherba, M., & Tarnavskyi, P. (2026). Rethinking memory hierarchy policies in DRAM+HBM systems. *Bulletin of Cherkasy State Technological University*, 31(2), 48-56. doi: 10.62660/bcstu/2.2026.48.

*Corresponding author (Maksym.Shcherba@lnu.edu.ua)

Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/)

of the hardware, leading to instances of sub-optimal resource utilisation. Consequently, there is a clear incentive to investigate refined management strategies that can more effectively align data placement with the actual bandwidth and latency profiles of modern heterogeneous memory systems.

The emergence of cache-coherent interconnects such as Compute Express Link (CXL) has recently attracted significant attention due to their capabilities in hardware heterogeneity management and resource disaggregation. D. Gouk *et al.* (2022) proposed DirectCXL, a system that connects a host processor complex directly to remote memory resources over the CXL memory protocol. Their research established that by bypassing data copies between host and remote memory, it is possible to expose the true performance of disaggregated resources to the user. Whilst such technologies facilitate hardware expansion, they also demand more sophisticated monitoring frameworks to handle the resulting complexity. S. Park *et al.* (2022) addressed this by introducing DAOS, an open-source system for data access-aware memory management. By utilising practical trade-offs between monitoring overhead and accuracy, DAOS provides a runtime system that auto-tunes management schemes for user-defined objectives, demonstrating that fine-grained access monitoring is essential for efficient placement decisions.

As memory systems scale to multi-terabyte capacities, the challenges of profiling and migration become even more pronounced. J. Ren *et al.* (2024) developed MTM, a framework for multi-tiered large memory that connects profiling overhead control with high-quality monitoring mechanisms. Their work indicates that a universal page migration policy, combined with huge page awareness, can significantly improve performance across realistic big-data workloads. Building upon these migration efficiencies, L. Xiang *et al.* (2024) introduced Nomad, which challenges the traditional exclusive tiering model. By implementing non-exclusive tiering through transactional page migration and page shadowing, Nomad mitigates memory thrashing and removes data movement from the critical execution path, significantly improving performance under memory pressure compared to standard transparent placement strategies. Beyond general page-level management, some investigators have identified the importance of specific system structures. S. Kannan *et al.* (2021) noted that most existing systems overlook kernel-level structures. Through the development of the KLOCs mechanism, they established that tiering kernel-level objects, such as inodes and socket buffers, significantly improves the efficiency of heterogeneous systems.

Alternative approaches focus on the granularity of management and the timing of data placement. The problem of flexible object-level management was addressed by B. Kammerdiener *et al.* (2023), who developed the bkmalloc method. They concluded that leveraging object-level information at the programming language level allows for more accurate data migration decisions between tiers compared to transparent OS-level mechanisms. The shortcomings of such transparent strategies were further detailed by D. Moura *et al.* (2022), who demonstrated that the standard AutoNUMA balancer yielded modest results for graph analytics due to a lack of temporal locality in data access patterns. Their characterisation showed that an object-level tiering approach could reduce execution times by an average of 21% by more accurately capturing application-specific behaviours than standard kernel heuristics. Similarly, M. B. Olson *et al.* (2022) focused on developing an online application guidance system, concluding that dynamically assigning application data based on live usage patterns achieves comparable performance to offline profiling methods while significantly outperforming standard unguided approaches. These developments suggest a trend toward more autonomous and context-aware systems that reduce the burden on the programmer whilst improving hardware efficiency.

Despite this significant number of developments, a critical gap remains in the current literature. Most analysed solutions are based on the assumption that DRAM is the fast tier compared to slower tiers like non-volatile memory or CXL expansion devices. Aspects of system functioning where DRAM coexists with higher-speed HBM, as well as questions of migration parameter calibration for such configurations, remain insufficiently studied. Existing policies are often motivated by capacity-driven assumptions, whereas for DRAM+HBM systems, the optimal tier designation may differ significantly for certain workloads. Furthermore, online page placement and migration without the necessity for offline profiling, pre-runs, or application recompilation are discussed only to a limited extent. Therefore, this work aimed to experimentally evaluate online tiered-memory management in DRAM+HBM systems without the requirement for offline preconditions. By investigating these dynamics, the study sought to determine how software-defined hierarchy policies could be reconciled with the physical bandwidth advantages of HBM, whilst identifying the efficiency thresholds where active management might become counter-productive.

MATERIALS AND METHODS

A modified memory management strategy based on the Ambix framework (Marques, 2021; Póvoas *et al.*, 2025) was proposed, and its performance was evaluated experimentally. The selection of Ambix was justified by its implementation as a Linux kernel module, which enabled the deployment of new policies without requiring modifications to the operating system kernel or application source code. The approach involved reconfiguring the memory hierarchy to designate HBM as the primary fast tier, whilst DRAM served as the secondary tier for less frequently accessed data. The research was based on a hypothesis that directly contradicted the standard assumptions of the Linux AutoNUMA

mechanism, where DRAM was viewed by default as the fastest hierarchy tier-a premise typically valid for DRAM+NVM or DRAM+CXL systems. Shifting priorities in favour of HBM was driven by its significant bandwidth advantage, which was a critical performance factor for data-intensive high-performance workloads. To optimise system performance, the page table scanning frequency was reduced to lower monitoring overhead, while per-cycle migration limits were increased to facilitate the rapid promotion of active data to the HBM tier. The experimental environment comprised a dual-socket server equipped with an Intel Xeon CPU Max 9462 processor featuring 32 cores per socket. The platform was configured with 256 GB of DRAM and 64 GB of HBM. All benchmark executions were restricted to a single CPU socket. The software stack utilised Rocky Linux 9.5 with Linux kernel version 6.1.

Performance evaluation was conducted using the HPCG (Dongarra *et al.*, 2016) and AMG Problem 2 (Yang *et al.*, 2017) benchmarks. For each run, it was calculated the standard performance metrics reported by the benchmarks as the Figure of Merit (FoM): HPCG performance in GFLOP/s and AMG performance as its reported FoM, which represents the rate of useful sparse-matrix work (i.e., effective matrix operations per second). To verify that the selected test scenarios were indeed bandwidth-bounded, a comparative analysis of results was performed in configurations where the entire memory volume was allocated exclusively in DRAM or exclusively in HBM. These tools provided standardised metrics to analyse system behaviour under bandwidth-intensive conditions.

To investigate the impact of resource availability, a memory offlining mechanism was employed to create controlled DRAM:HBM capacity ratios of 2:1, 1:1, and 1:2. Memory offlining was performed using the Linux memory hotplug interface via sysfs, which allows selected memory blocks to be disabled to enforce the desired ratios. The resulting data were recorded across six distinct system configurations, enabling a comparison between the proposed solution and standard kernel mechanisms, specifically the AutoNUMA balancing system (Huang, 2019). This configuration ensured that the benchmarks occupied nearly all available memory within a single socket, providing a comprehensive understanding of how different policies influenced computational efficiency.

RESULTS AND DISCUSSION

Initial performance comparisons demonstrated that allocating the entire memory volume to HBM yielded substantial throughput improvements over configurations utilising only DRAM. As shown in Figure 1, data placement within HBM provided a 24% performance gain for the HPCG benchmark and a 22% gain for the AMG benchmark. The observed performance gap between DRAM-only and HBM-only configurations validated the bandwidth-intensive nature of the workloads.

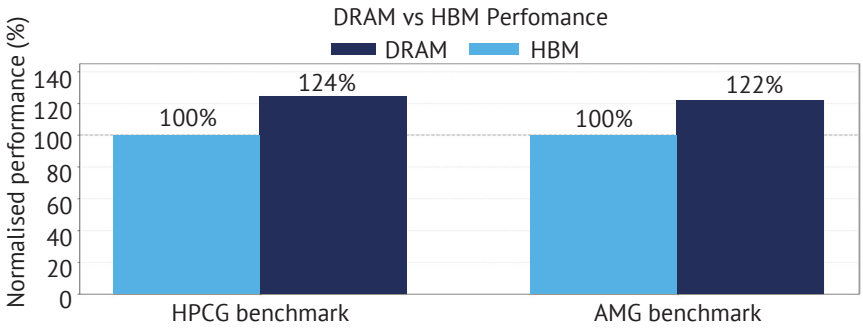


Figure 1. Performance comparison of DRAM and HBM memory across benchmarks
Source: created by the authors

Assessments across varying DRAM:HBM capacity ratios indicated that standard operating system mechanisms could lead to sub-optimal resource utilisation. For the HPCG workload, the *No balancing* baseline yielded performance between 33.62 and 34.93 GFLOP/s, as summarised in Table 1. The activation of NUMA balancing resulted in a performance degradation of up to 4% relative to the baseline – a decrease attributed to the misalignment of page promotion directions, where the kernel incorrectly identified the slower DRAM tier as the fast tier and promoted pages accordingly.

Table 1. HPCG benchmark performance (GFLOP/s) across different DRAM:HBM capacity ratios and memory management policies

Configuration	Performance (GFLOP/s)		
	DRAM:HBM = 2:1	DRAM:HBM = 1:1	DRAM:HBM = 1:2
No balancing	33.62	33.85	34.93
NUMA balancing	32.44	32.40	33.97
Ambix (DRAM-priority)	29.06	29.63	30.87

Continued Table 1.

Configuration	Performance (GFLOP/s)		
	DRAM:HBM = 2:1	DRAM:HBM = 1:1	DRAM:HBM = 1:2
Ambix (HBM-priority)	31.14	30.05	31.42
Ambix (DRAM-priority, tuned)	29.42	31.46	34.19
Ambix (HBM-priority, tuned)	35.86	35.38	36.05

Source: created by the authors

Performance metrics for the HPCG benchmark, illustrated in Figure 2, revealed significant variations based on the management policy employed. The original Ambix (DRAM-priority) configuration resulted in a performance decline of 12-14% relative to the baseline. It was observed that implementing individual modifications to this framework yielded incremental improvements, though neither was sufficient to outperform the baseline independently. Reconfiguring the hierarchy to prioritise HBM – Ambix (HBM-priority) – partially mitigated the loss, reducing degradation to 7-11%. Similarly, maintaining the DRAM-priority while optimising internal framework logic – Ambix (DRAM-priority, tuned) – reduced the performance penalty to 2-12%.

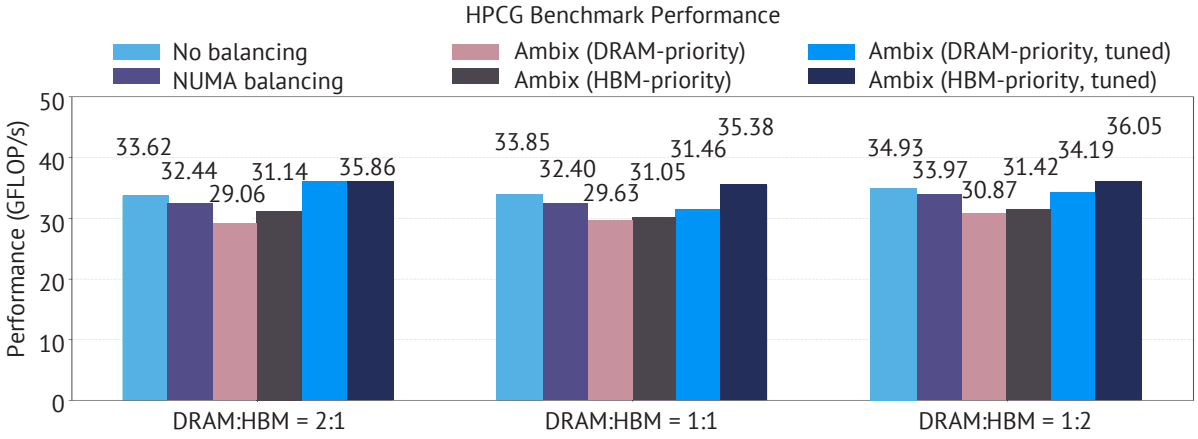


Figure 2. Comparison of HPCG benchmark performance (GFLOP/s) across evaluated memory management policies

Source: created by the authors

Net performance gains were only realised when both modifications were applied simultaneously. Under the *Ambix (HBM-priority, tuned)* framework, which combined an HBM-fast hierarchy with optimised scan frequencies and migration limits, performance improvements of 7%, 5%, and 3% were achieved for the 2:1, 1:1, and 1:2 DRAM:HBM ratios, respectively. These findings indicated that while each structural change provided a marginal benefit, the synergy between a hardware-aligned hierarchy and tuned migration parameters was essential to surpass the performance of the baseline.

Furthermore, the results suggested that the effectiveness of proactive memory placement diminished as the availability of HBM increased. At lower HBM capacities, efficient data placement was critical to maximise throughput, whereas higher HBM availability allowed a larger portion of the hot data to reside in the fast tier without intervention, thereby diminishing the relative impact of managed migration. Performance metrics for the AMG benchmark were summarised in Table 2, where the baseline provided a FoM ranging from $1.25 \cdot 10^9$ to $1.36 \cdot 10^9$. In contrast to the results observed for HPCG, the activation of *NUMA balancing* yielded performance identical to the baseline, indicating that the default kernel mechanism neither improved nor degraded throughput for this specific workload. However, the broader results for managed configurations demonstrated a trend consistent with the overhead typically associated with active memory management in tiered systems.

Table 2. AMG benchmark performance (FoM) across different DRAM:HBM capacity ratios and memory management policies

Configuration	Figure of Merit		
	DRAM:HBM = 2:1	DRAM:HBM = 1:1	DRAM:HBM = 1:2
No balancing	$1.25 \cdot 10^9$	$1.26 \cdot 10^9$	$1.36 \cdot 10^9$
NUMA balancing	$1.25 \cdot 10^9$	$1.26 \cdot 10^9$	$1.36 \cdot 10^9$
Ambix (DRAM-priority)	$0.99 \cdot 10^9$	$1.00 \cdot 10^9$	$1.04 \cdot 10^9$

Continued Table 2.

Configuration	Figure of Merit		
	DRAM:HBM = 2:1	DRAM:HBM = 1:1	DRAM:HBM = 1:2
Ambix (HBM-priority)	$1.06 \cdot 10^9$	$1.07 \cdot 10^9$	$1.10 \cdot 10^9$
Ambix (DRAM-priority, tuned)	$1.19 \cdot 10^9$	$1.21 \cdot 10^9$	$1.29 \cdot 10^9$
Ambix (HBM-priority, tuned)	$1.27 \cdot 10^9$	$1.27 \cdot 10^9$	$1.34 \cdot 10^9$

Notes: the Figure of Merit is the performance metric reported by the AMG benchmark
Source: created by the authors

The introduction of managed placement via the Ambix (DRAM-priority) policy resulted in a significant performance reduction of 21-23% relative to the baseline. Incremental improvements were observed when modifications were applied individually to the framework; these trends were visualised for comparative analysis in Figure 3 to highlight the relative impact of each change. Prioritising the high-bandwidth tier in the Ambix (HBM-priority) configuration reduced the performance penalty to 15-19% compared to the baseline. Similarly, maintaining the original hierarchy while applying parameter optimisations in the Ambix (DRAM-priority, tuned) configuration further limited the degradation to 4-5%.

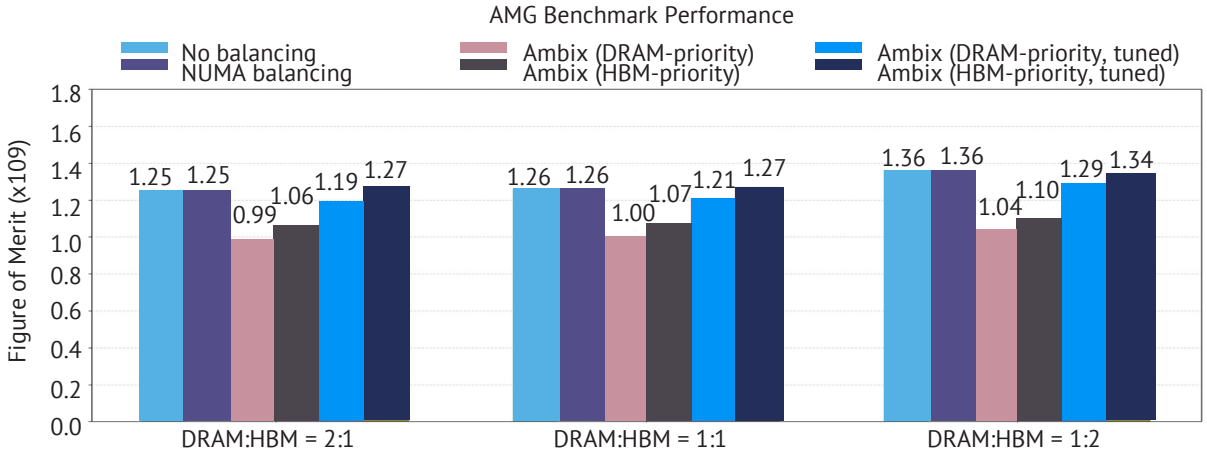


Figure 3. Comparison of AMG benchmark performance (Figure of Merit) across evaluated memory management policies
Notes: the Figure of Merit is the performance metric reported by the AMG benchmark
Source: created by the authors

As illustrated in Figure 3, the highest efficiency across all evaluated configurations was achieved through the simultaneous application of hierarchy reconfiguration and parameter tuning. Under the Ambix (HBM-priority, tuned) policy, the system showed a 1.4% improvement over the baseline at a 2:1 DRAM:HBM capacity ratio, although a 1.5% degradation was experienced at the 1:2 ratio. This slight decline at higher HBM capacities was attributed to the variable nature of the AMG memory footprint. At the onset of the workload, the active dataset could fit entirely within the HBM tier. In such instances, the computational overhead of the management framework outweighed the benefits of page migration, rendering active management counter-productive when HBM resources were sufficiently abundant to accommodate the workload without intervention. In summary, the evaluation across both benchmarks underscored that effective tiered memory management in DRAM+HBM systems required a synergistic approach. While individual modifications – such as aligning the memory hierarchy with hardware bandwidth in Ambix (HBM-priority) or optimising migration logic in Ambix (DRAM-priority, tuned) – yielded incremental benefits, neither was sufficient to consistently surpass the No balancing baseline. Net performance improvements were only realised through the unified Ambix (HBM-priority, tuned) policy, which reconciled the software-defined hierarchy with the physical memory topology. Furthermore, the results highlighted a critical efficiency threshold: as the availability of HBM increased relative to the workload footprint, the computational overhead of the management framework began to outweigh the benefits of proactive migration. This phenomenon was particularly evident in the AMG results, suggesting that the utility of managed memory placement was most pronounced in resource-constrained scenarios where the high-bandwidth tier could not accommodate the entire active dataset without intervention.

The observed performance dynamics in DRAM+HBM systems indicated a fundamental departure from the capacity-driven paradigms that dominated tiered memory research over the last decade. While conventional

heterogeneous memory management had focused on expanding DRAM capacity through slower tiers like CXL, the integration of HBM inverted the performance hierarchy. In this bandwidth-driven context, the secondary tier was significantly faster than the primary capacity tier. The recorded 4% performance degradation under the NUMA balancing configuration underscored the failure of legacy locality heuristics in such environments. As highlighted by H. Al Maruf *et al.* (2023), the Linux kernel was originally designed for homogeneous architectures and traditionally prioritised memory nodes based on relative NUMA distances, which fail to reflect the actual bandwidth capabilities of heterogeneous tiers. Consequently, if a thread was perceived as local to the DRAM node, standard mechanisms inadvertently promoted pages to the slower tier, a finding that aligned with the research of J. Kim *et al.* (2021) regarding multi-tiered systems. This study confirmed that tier-agnostic policies, which prioritised proximity over throughput, were fundamentally ill-equipped for asymmetric bandwidth hierarchies.

The limitations of such heuristics extended beyond standard kernel mechanisms, manifesting even more severely in specialised management frameworks, as evidenced by the significant performance reduction of 21-23% observed in the Ambix (DRAM-priority) configuration. This provided empirical evidence that software assumptions must align with hardware realities, yet merely reconfiguring the hierarchy to prioritise the high-bandwidth tier (Ambix (HBM-priority)) proved insufficient, as it only reduced the performance penalty to 15-19%. This residual gap suggests that a correct hierarchy is only the first requirement; the second essential pillar is the precise calibration of management logic. Similar conclusions were reported by M. Kim *et al.* (2025) for tiered-memory management under co-located workloads, where correct tier preference must be complemented by adaptive runtime control. The sensitivity of tiered systems to such logic was validated by the work of T.D. Doudali *et al.* (2021), whose Cori framework demonstrated that sub-optimal migration frequencies could lead to a 10-100% performance degradation. These findings explain why surpassing the baseline in this study required a synergy between an HBM-priority hierarchy and carefully tuned migration limits. As J.D. McCalpin (2023) established, sustained bandwidth in Xeon Max processors is already constrained by hardware limiters such as insufficient memory concurrency, making the system exceptionally vulnerable to software overheads that arise from uncalibrated management policies. The most pervasive of these overheads originated from the methodological challenges associated with kernel-level monitoring remained a primary friction point. A. Raybuck *et al.* (2021) discovered in their work on HeMem that page table scanning incurred prohibitive costs due to TLB shootdowns and lock contention. While this study utilised a kernel-integrated approach, the performance gains achieved by reducing scanning frequency in the Ambix (HBM-priority, tuned) configuration served as a necessary concession to these inherent costs. This strategy was consistent with the research of T. Lee *et al.* (2023), who established in their work on MEMTIS that fixed-rate scanning mechanisms faced a steep overhead-accuracy trade-off, with CPU costs reaching 72.85% to maintain high monitoring precision. They demonstrated that an effective management layer required dynamic page classification via multiple access thresholds, which enabled a significant reduction in sampling intensity for colder memory regions to mitigate system-wide overhead. Such optimisations were crucial, as the sensitivity of the HBM-DRAM gap made management overhead even more critical than in latency-heavy NVM systems. Recent work such as FlexMem (Xu *et al.*, 2024) supports this point: by adapting profiling and demotion decisions, it avoids scanning with fixed ratio and reduces unnecessary page migrations in tiered memory.

The tangible impact of these overheads was most clearly manifested in the identification of an efficiency threshold, illustrated by the AMG benchmark results. The transition from a 1.4% improvement at a 2:1 ratio to a 1.5% degradation at a 1:2 ratio suggested that active memory management possessed a point of diminishing returns. When HBM resources were sufficiently abundant to accommodate the active dataset at the onset of a workload, the computational overhead associated with page table scanning outweighed any potential performance benefits. Ultimately, the comparative analysis suggested that while industry efforts were largely focused on masking the slowness of expansion tiers, HBM required a paradigm that exploited primary tier bandwidth while minimising management interference. As explored by P. Duraisamy *et al.* (2023) in the TMTS framework, hardware-guided tiering often scaled better than software-only scanning. This study validated that an HBM-priority policy acted as a necessary proxy for bandwidth-priority, forcing data into the wider pipe even when latency signals were ambiguous. The synergy between hierarchy prioritisation and parameter tuning was therefore a fundamental requirement for unlocking the performance potential of next-generation heterogeneous memory systems.

CONCLUSIONS

This study demonstrated that effective utilisation of heterogeneous memory systems, particularly for DRAM+HBM configurations, required a well-defined and carefully considered memory hierarchy. The primary outcome of this research was the confirmation that reconsidering the memory hierarchy, by treating HBM as a top-tier memory in a multi-level system, could significantly enhance performance for bandwidth-dependent workloads. This finding challenged the default assumptions embedded in most modern operating

systems, where DRAM was typically regarded as the fastest memory and thus prioritised for page promotion. The proposed modification, based on the Ambix system, achieved improved efficiency through two key mechanisms: preferential use of HBM and fine-tuning of internal page migration parameters. Experimental evaluations using standard benchmarks revealed consistent performance improvements across most tested configurations. The most substantial gains were observed in scenarios with limited HBM capacity, where intelligent data placement became critical.

In contrast, when HBM capacity was sufficient to store the full active data used in a benchmark, the benefits of controlled page migration lessened or might even have become negative due to overhead costs. This highlighted the importance of adaptive approaches that considered memory utilisation and workload characteristics. In systems with limited DRAM capacity where HBM was added as a supplementary but also constrained memory tier, the proposed approach significantly improved performance by optimising data placement and memory bandwidth usage. Overall, the results indicated that an appropriate memory policy could improve system efficiency without requiring modifications to application code or operating system kernel-an advantage that facilitated deployment especially in large-scale systems. Future work could focus on evaluating and adapting alternative heterogeneous-memory managers to clarify the practical differences between HBM- and DRAM-prioritisation strategies. In addition, a promising research direction is the design and implementation of a workload-aware runtime memory manager that dynamically determines the most suitable memory tier for each page and adapts monitoring intensity based on lightweight indicators such as bandwidth demand, CPU load, and working-set residency across all memory tiers.

ACKNOWLEDGEMENTS

We are grateful to Jakub Schmiegel for his valuable technical assistance and to Łukasz Kucharski for insightful comments that significantly improved the manuscript. We also thank both for their constructive discussions regarding the results.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Dongarra, J., Heroux, M.A., & Luszczek, P. (2016). High-performance conjugate-gradient benchmark: A new metric for ranking high-performance computing systems. *The International Journal of High Performance Computing Applications*, 30(1), 3-10. doi: 10.1177/10943420155.

[2] Doudali, T.D., Zahka, D., & Gavrilovska, A. (2021). Cori: Dancing to the right beat of periodic data movements over hybrid memory systems. *2021 IEEE international parallel and distributed processing symposium (IPDPS)* (pp. 350-359). Piscataway: IEEE. doi: 10.1109/IPDPS49936.2021.00043.

[3] Duraisamy, P., *et al.* (2023). Towards an adaptable systems architecture for memory tiering at warehouse-scale. In *Proceedings of the 28th ACM international conference on architectural support for programming languages and operating systems* (pp. 727-741). New York: Association for Computing Machinery. doi: 10.1145/3582016.3582031.

[4] Gouk, D., Lee, S., Kwon, M., & Jung, M. (2022). Direct access, high-performance memory disaggregation with DirectCXL. In *2022 USENIX annual technical conference (USENIX ATC 22)* (pp. 287-294). Carlsbad: USENIX Association.

[5] Al Maruf, H., Wang, H., Dhanotia, A., Weiner, J., Agarwal, N., Bhattacharya, P., Petersen, C., Chowdhury, M., Kanaujia, S., & Chauhan, P. (2023). TPP: Transparent page placement for CXL-enabled tiered-memory. In *Proceedings of the 28th ACM international conference on architectural support for programming languages and operating systems* (pp. 742-755). New York: Association for Computing Machinery. doi: 10.1145/3582016.3582063.

[6] Huang, Y. (2019). *Autonuma: Optimize memory placement in memory tiering system*. Retrieved from https://lkml.org/lkml/2021/2/4/289.

[7] Kammerdiener, B., McMichael, J.Z., Jantz, M.R., Doshi, K.A., & Jones, T. (2023). Flexible and effective object tiering for heterogeneous memory systems. In *Proceedings of the 2023 ACM SIGPLAN international symposium on memory management* (pp. 163-175). New York: Association for Computing Machinery. doi: 10.1145/3591195.3595277.

[8] Kannan, S., Ren, Y., & Bhattacharjee, A. (2021). KLOCs: Kernel-level object contexts for heterogeneous memory systems. In *Proceedings of the 26th ACM international conference on architectural support for programming languages and operating systems* (pp. 65-78). New York: Association for Computing Machinery. doi: 10.1145/3445814.3446745.

[9] Kim M., Han, S., Park, G., & Kim, D. (2025). MTAT: Adaptive fast memory management for co-located latency-critical workloads in tiered memory system. In *Proceedings of the 26th international middleware conference* (pp. 86-98). New York: Association for Computing Machinery. doi: 10.1145/3721462.3770767.

[10] Kim, J., Choe, W., & Ahn, J. (2021). Exploring the design space of page management for multi-tiered memory systems. In *2021 USENIX annual technical conference (USENIX ATC 21)* (pp. 715-728). Berkeley: USENIX Association.

[11] Lee, T., Monga, S., Min, C., & Eom, Y.I. (2023). MEMTIS: Efficient memory tiering with dynamic page classification and page size determination. In *Proceedings of the 29th ACM symposium on operating systems principles* (pp. 17-34). doi: 10.1145/3600006.3613167.

[12] Marques, M.S. (2021). *Ambix: Rethinking Linux's page management to support the new Intel Optane DC persistent memory*. (Master's thesis, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal).

[13] McCalpin, J.D. (2023). Bandwidth limits in the Intel Xeon Max (sapphire rapids with HBM) processors. In *High performance computing: ISC high performance 2023 international workshops* (pp. 403-413). Berlin: Springer-Verlag. doi: 10.1007/978-3-031-40843-4_30.

[14] Moura, D., Mosse, D., & Petrucci, V. (2022). Performance characterization of AutoNUMA memory tiering on graph analytics. In *2022 IEEE international symposium on workload characterization (IISWC)* (pp. 171-184). IEEE. doi: 10.1109/iiswc55918.2022.00024.

[15] Olson, M.B., Kammerdiener, B., Jantz, M.R., Doshi, K.A., & Jones, T. (2022). Online application guidance for heterogeneous memory systems. *ACM Transactions on Architecture and Code Optimization*, 19(3), article number 45. doi: 10.1145/3533855.

[16] Park, S., Bhowmik, M., & Uta, A. (2022). DAOS: Data access-aware operating system. In *Proceedings of the 31st international symposium on high-performance parallel and distributed computing* (pp. 4-15). New York: Association for Computing Machinery. doi: 10.1145/3502181.3531466.

[17] Póvoas, J., Barreto, J., Chomiski, B., Gonçalves, A., Karabeinikau, F., Maciejewski, M., Schmiegel, J., & Storozhuk, K. (2025). Better memory tiering, right from the first placement. In *Proceedings of the 16th ACM/SPEC international conference on performance engineering* (pp. 253-265). New York: Association for Computing Machinery. doi: 10.1145/3676151.3719378.

[18] Raybuck, A., Stamler, T., Zhang, W., Erez, M., & Peter, S. (2021). HeMem: Scalable tiered memory management for big data applications and real NVM. In *Proceedings of the ACM SIGOPS 28th symposium on operating systems principles* (pp. 392-407). New York: Association for Computing Machinery. doi: 10.1145/3477132.3483550.

[19] Ren, J., Xu, D., Ryu, J., Shin, K., Kim, D., & Li, D. (2024). MTM: Rethinking memory profiling and migration for multi-tiered large memory. In *Proceedings of the nineteenth european conference on computer systems* (pp. 803-817). New York: Association for Computing Machinery. doi: 10.1145/3627703.3650075.

[20] Xiang, L., Lin, Z., Deng, W., Lu, H., Rao, J., Yuan, Y., & Wang, R. (2024). Nomad: Non-exclusive memory tiering via transactional page migration. In *18th USENIX symposium on operating systems design and implementation (OSDI 24)* (pp. 19-35). Santa Clara: USENIX Association.

[21] Xu, D., Ryu, J., Shin, K., Su, P., & Li, D. (2024). FlexMem: Adaptive page profiling and migration for tiered memory. In *Proceedings of the 2024 USENIX annual technical conference (USENIX ATC 24)* (pp. 817-833). Santa Clara: USENIX Association.

[22] Yang, U., Falgout, R., & Park, J. (2017). *Algebraic multigrid benchmark*. Retrieved from https://www.osti.gov/biblio/1389816.

Переосмислення політик ієрархії пам’яті в системах DRAM+HBM

Максим Щерба
Аспірант
Львівський національний університет імені Івана Франка
79007, вул. Університетська, 1, м. Львів, Україна
https://orcid.org/0009-0008-0332-9254
Петро Тарнавський
Аспірант
Львівський національний університет імені Івана Франка
79007, вул. Університетська, 1, м. Львів, Україна
https://orcid.org/0000-0003-3265-8168

Анотація. Ефективне управління різнорідними системами пам’яті, що містять пам’ять із високою пропускною здатністю (HBM), є критично важливим для загальної продуктивності. Традиційні підходи часто припускають, що динамічна пам’ять з довільним доступом (DRAM) є найшвидшим рівнем, що може не справджуватися в конфігураціях з HBM. Метою цього дослідження було оцінювання впливу на продуктивність переналаштування політики ієрархії пам’яті шляхом зміни переваги з DRAM на HBM у системах DRAM + HBM. У дослідженні використано модифіковану систему Ambix, у якій адаптовано політики міграції для призначення HBM швидким рівнем та налаштовано внутрішні параметри для кращої відповідності архітектурі DRAM + HBM. Оцінювання проводилося за допомогою тестів HPCG та AMG на двопроцесорному сервері. Встановлено, що розміщення всіх даних у пам’яті HBM забезпечило приріст продуктивності до 24 % порівняно з використанням лише DRAM. Аналіз продемонстрував, що стандартне балансування NUMA в Linux могло спричинити деградацію продуктивності до 4 %, оскільки воно некоректно переміщувало сторінки пам’яті до повільнішого рівня DRAM. Дослідження виявило, що хоча стандартні конфігурації Ambix могли знижувати продуктивність до 23 %, надання переваги HBM покращило результати на величину від 1,4 % до 22 % порівняно з політиками, орієнтованими на DRAM. Крім того, налаштовано внутрішні параметри, що дозволило зменшити частоту сканування таблиць сторінок та збільшити ліміти міграції за цикл. Ці модифікації забезпечили приріст продуктивності від 3 % до 7 % для тесту HPCG порівняно з базовою системою. Для тесту AMG експериментальні результати показали варіювання від деградації на 1,5 % до покращення на 1,4 %, залежно від співвідношення обсягів DRAM:HBM. Отримані результати дозволяють системним архітекторам оптимізувати ієрархію пам’яті в серверах наступного покоління без необхідності внесення змін до ядра ОС або коду програм

Ключові слова: різнорідні системи пам’яті; управління пам’яттю; динамічна пам’ять з довільним доступом; пам’ять з високою пропускною здатністю; ієрархія пам’яті

BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 2. 2026

UDC 004.413.4:004.65
DOI: 10.62660/bcstu/2.2026.57

ISSN: 2306-4412
E-ISSN: 2708-6070

Architectural and methodological approaches to the design of corporate and web-oriented information systems with elements of intelligent data analysis

Vladyslav Kozub*
PhD in Computer Science, Associate Professor
Luhansk Taras Shevchenko National University
37500, 2 Viktor Novikov Str., Lubny, Ukraine
https://orcid.org/0000-0003-2710-7206

Olha Smahina
PhD in Pedagogical Sciences, Associate Professor
Luhansk Taras Shevchenko National University
37500, 2 Viktor Novikov Str., Lubny, Ukraine
https://orcid.org/0000-0002-6024-5152

Halyna Kozub
PhD in Technical Sciences, Associate Professor
Luhansk Taras Shevchenko National University
37500, 2 Viktor Novikov Str., Lubny, Ukraine
https://orcid.org/0000-0001-5387-050X

Abstract. The relevance of the study is due to the digital transformation and the need to improve architectural and methodological approaches to the design of intelligent corporate and web-oriented information systems. The purpose of the study was to form a conceptual approach to the design of information systems using intelligent data analysis tools. The study was based on the theoretical analysis of scientific materials and the application of systemic, conceptual, project-oriented, domain-oriented and data-driven methodological approaches to substantiate the principles of designing adaptive information systems. The results obtained demonstrated the transition of modern information systems to multi-level modular and data-centric architectures that combine functional flexibility, scalability and intelligent analytical data processing. The integration of artificial intelligence algorithms provides the implementation of predictive analytics, information classification and detection of anomalies in the system's behaviour. The effectiveness of the platforms is assessed through a complex of time, resource and analytical metrics, in particular latency, throughput, central processing unit utilisation, memory consumption, error rate, mean absolute error, root mean squared error, mean absolute percentage error, precision, and recall. It is shown that the reliability of information systems is ensured by the use of fault-tolerant architectures, service component redundancy and self-tuning computing environments. The generalisation of the results confirmed the prospects for the development of adaptive cognitive-oriented data-centric information platforms of the new generation. The results can be used by scientists, developers, and software architects and organisations involved in the digital transformation of information systems in the design, optimisation and implementation of adaptive analytically oriented digital platforms

Keywords: machine learning; big data; data-driven; scalability; adaptive information platforms; domain-oriented design

Article's History: Received: 17.11.2025; Revised: 16.03.2026; Accepted: 18.05.2026; Published: 26.06.2026

Suggested Citation:
Kozub, V., Smahina, O., & Kozub, H. (2026). Architectural and methodological approaches to the design of corporate and web-oriented information systems with elements of intelligent data analysis. *Bulletin of Cherkasy State Technological University*, 31(2), 57-72. doi: 10.62660/bcstu/2.2026.57.

*Corresponding author (VladyslavKozub052@outlook.com)

Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/)

INTRODUCTION

The relevance of the study is determined by the intensive digital transformation of society and the business environment, which necessitates the creation of more complex and functionally rich information systems. The growth of structured and unstructured data requires a transition from traditional transactional platforms to intelligent digital environments capable of providing automated analytical processing of information and decision-making support. In these conditions, the development of coordinated architectural and methodological approaches to the design of corporate and web-oriented information systems is used.

In the scientific discourse, information systems research is increasingly focused on issues of architectural optimisation of digital platforms and the integration of intelligent data processing methods into applied information environments. In the work of H. Khudov *et al*. (2025), an analysis of architectural models of information technologies for image processing in complex technical systems was carried out, where it was shown that the efficiency of the functioning of the information platform is determined by the structural organisation of data processing channels and the choice of algorithmic procedures for transforming information flows. The authors proved that multi-level data representation provides increased accuracy of signal processing and improves the stability of monitoring systems. In the study of V. Obukhov & V. Tovbych (2025), the emphasis is placed on the application of intelligent technologies in the processes of architectural design of information systems. The authors emphasised that the use of algorithmic methods of design support makes it possible to formalise the adoption of architectural decisions and reduce the subjectivity of engineering analysis. The results obtained showed that the integration of artificial intelligence methods into design processes contributes to increasing the adaptability of architectural models and optimising the structural complexity of software platforms.

Another perspective is revealed by J. Luo *et al*. (2022), who studied the application of data warehouse technologies and methods of intelligent information analysis in corporate information systems. The authors showed that the combination of the concepts of data warehouse and data mining provides effective support for analytical processing of large volumes of information, in particular in the tasks of business process management and forecasting system behaviour. The authors emphasised that the use of structured analytical environments increases the efficiency of decision-making and contributes to the formation of integrated digital platforms. The work of T. Rak & R. Żyła (2022) demonstrated the possibilities of using data mining to detect hidden dependencies in the output streams of web systems. The authors showed that the use of analytical procedures allows identifying structural and behavioural patterns of the functioning of the information environment, which is important for the tasks of optimising the processing of user requests and improving the quality-of-service interaction. The issue of architectural analysis of corporate information systems based on an application programming interface (API) gateway logs is considered in the study of C.R. Pinheiro *et al*. (2024). The authors proposed the use of a lightweight ontological model for interpreting architectural patterns of service interaction, which provides the possibility of formalised analysis of the structure of the corporate platform and monitoring its functional behaviour. The results obtained demonstrated the promising nature of combining semantic modelling and telemetric analysis of system events.

Another area of research is presented in the work of M. Muslih *et al*. (2024), where a web-based information system for assessing professional suitability of users was developed using the architectural principles of the Enterprise Architecture Framework. The authors proved that the use of structured methodological approaches to the design of digital platforms contributes to increasing the coherence of business processes, technological infrastructure and analytical services of the information system. The work of J. Vinueza-Martinez *et al*. (2024) investigated modern trends in the development of geographic information systems built on the basis of the web geographic information system (GIS) architecture. The authors conducted a bibliometric analysis of scientific publications and showed that the development of web-based geospatial platforms is characterised by increasing attention to the integration of spatial analytics, cloud computing and service-oriented models of geodata processing. The results of the study confirmed that modern GIS platforms are transforming into multifunctional analytical environments capable of supporting the monitoring of territorial processes and making management decisions based on spatiotemporal data. At the same time, the study by L. Abdulrahman *et al*. (2022) presents an overview of modern approaches to the design of web-based regional information systems, in particular in the field of tourism management. The authors emphasised that the effectiveness of such platforms is determined by the comprehensive integration of functional services, analytical modules and user interfaces within a single digital architecture.

Previous studies mainly focused on the analysis of individual aspects of information systems design, in particular architectural models of corporate and web-oriented platforms or the application of data mining methods in specific application tasks. At the same time, the issues of system integration of architectural, methodological and analytical components within a single conceptual model of information systems design remain insufficiently addressed. In particular, there is a gap in substantiating the choice of architectural solutions in relation to data processing methods, assessing platform effectiveness and demonstrating

possible scenarios for the practical implementation of intelligent modules in corporate and web-oriented environments. The aim of the study was to develop a conceptual approach to the design of corporate and web-oriented information systems with the integration of data mining elements. To achieve this goal, the following tasks were solved: to analyse modern architectural and methodological approaches to the design of information systems, taking into account the possibilities of integrating analytical modules; to substantiate the conceptual model of the information platform and determine approaches to assessing the impact of the proposed solutions on the scalability, adaptability, and performance of the system.

MATERIALS AND METHODS

The study was of a theoretical and survey nature, with project-oriented and analytical approaches. The time frame of the study covered the period from September 2024 to February 2026. The materials of the study were scientific articles and reviews on the architecture of information systems and digital transformation, as well as materials from international conferences on distributed computing and big data (Adamov, 2018; Zhang, 2025). To analyse the architecture of web-oriented platforms, the following standards and technical documents were used: OpenAPI Specification 3.1.0, Request for Comments (RFC) 7231 (HyperText Transfer Protocol (HTTP) semantics), RFC 3986 (Uniform Resource Identifier (URI) syntax), W3C Web API Best Practices recommendations, JavaScript Object Notation (JSON) Schema specifications, as well as OAuth 2.0 security standards for describing API integration. The system analysis method was used to consider the information system as a complex multi-level object. Within the framework of this approach, the logical structuring of the digital platform, the definition of functional subsystems and the analysis of the relationships between the infrastructure, service, data-centric and analytical components of the architecture were carried out. System analysis made it possible to form a generalised concept of the functioning of modern information ecosystems and to determine the principles of structural integrity of the software platform.

The conceptual modelling method was used to formalise the architectural principles of building an information system and describe the general logic of its functioning. Within the framework of the study, a theoretical structural representation of the digital platform was carried out in the form of a generalised multi-level model, which included infrastructure, service, data-centric, analytical and application components. Special attention was paid to the analysis of the principles of organising interaction between the functional levels of the system and determining the general place of intellectual and analytical components in the structure of the information platform. The project-oriented method was used to study the theoretical aspects of the life cycle of an information system as an engineering-controlled process of creating software platforms. Within the framework of this approach, the general principles of organising the stages of information systems design were considered, including the formation of requirements for functionality, conceptual architectural design, implementation of software modules, testing of software logic, system implementation and subsequent support of the digital platform. Methodologically, the study was based on a generalised analysis of modern approaches to the modular organisation of software, the principles of weak component coupling and iterative logic of information systems development without performing applied experimental procedures.

The domain-oriented design method was used to ensure semantic consistency of the software architecture with the subject area of the system's functioning. Within the framework of the study, a business domain model was formed, in which the platform architecture was organised around key business objects. Data processing service modules and information repository storage mechanisms were defined, which allowed reducing the cognitive complexity of the architecture and ensuring the possibility of reusing business logic during system modernisation. Data-driven methodology was used to justify architectural solutions based on the analysis of the operational characteristics of the information platform. Based on the analysed data, a model of adaptive configuration management of the software environment was formed, which provided for automated optimisation of system parameters depending on changes in load and user behavioural characteristics.

The algorithmic component of the study was based on a theoretical analysis of the possibilities of applying data mining methods without performing empirical computational experiments. The study considered the conceptual principles of the functioning of forecasting models based on time series analysis and statistical learning, which allowed extrapolating the behavioural characteristics of information systems in hypothetical scenarios of the development. For the tasks of classification and segmentation of information objects, the methodological features of the application of clustering algorithms, discriminant analysis and deep learning methods were analysed from the standpoint of the functional purpose in the structure of the information platform. The study used the method of comparative analysis of architectural models of corporate information systems. The analysis was carried out according to key technical criteria, in particular, scalability, modularity, complexity of support, independence of components and the level of fault tolerance, which made it possible to generalise the features of monolithic, service-oriented and microservice architectures. An architectural analysis of web-oriented systems was conducted, in particular, the principles of frontend/backend decomposition, API-oriented interaction, and the Representational State Transfer (REST) approach were

investigated. A structural and functional analysis of data integration and processing was also used, which covered the concepts of Data warehouse, Extract, Transform, Load, Extract, Load, Transform, and visualisation tools. In addition, an analysis of event-oriented interaction of components in distributed information systems was carried out to assess the mechanisms of coordination and data transfer between services. A historical and evolutionary analysis was also applied, which was used to trace the stages of development of information system architectures and approaches to the design – from monolithic systems to service-oriented and microservice architectures, as well as to analyse the transformation of data integration methods and the organisation of component interaction in modern digital platforms.

The analysis of the reliability of the information system was carried out from the standpoint of the concept of fault-tolerant architectural environments at the level of theoretical modelling of possible scenarios of the platform's operation. The principles of service component redundancy, automated recovery of programme processes, and isolation of critical infrastructure nodes as mechanisms for increasing the stability of a digital system were considered. The study of the adaptability of the information platform was based on the concept of self-tuning computing environments, in which the parameters of the software infrastructure could change in accordance with variations in the operational load or behavioural characteristics of users based on the analysis of telemetric data in theoretically simulated conditions. An integral assessment of the effectiveness of the information system architecture was carried out at the level of conceptual synthesis of individual groups of indicators without conducting numerical experiments. The combined performance index was considered as a generalised theoretical construct that reflected the balance between speed, resource efficiency, accuracy of analytical models, and platform stability. Thus, the set of applied materials and methods made it possible to form a holistic theoretical and analytical substantiation of the principles of designing modern adaptive information systems, providing a systematic analysis of architectural models, methodological approaches, and concepts for integrating intelligent analytics into digital platforms.

RESULTS

Conceptual principles for building corporate and web-based information systems

The 2016-2026 stage of information technology development is characterised by a systemic transformation of approaches to designing information platforms, which is due to the transition to digitally-oriented data and process management models. The evolution of corporate and web-based information systems is directly related to the concept of digital transformation, which involves the integration of computing, communication and analytical components into a single technological ecosystem. Within this paradigm, the information system is considered not only as a tool for automating operational activities, but also as an intelligent decision-making support platform capable of adapting to changes in the external and internal environment. The evolutionary development of information systems architecture went through several conceptual stages, ranging from monolithic software complexes to distributed service-oriented platforms. Classic monolithic solutions were characterised by a high level of internal connectivity of components and limited scalability, which made it difficult to modernise the system in the face of growing data volumes. Further development led to the formation of the Service-Oriented Architecture paradigm, which involves the decomposition of functionality into independent service modules with standardised interaction interfaces. This approach provides the possibility of reusing services, increases the level of integration compatibility and simplifies software maintenance processes. The next logical step was the concept of Microservices architecture, which considers the system as a set of autonomous microcomponents, each of which implements a separate business function and can be deployed independently of other modules (Efe, 2024; Molodets & Bulana, 2024).

The digital transformation of information systems is largely determined by the exponential growth in the volume of structured and unstructured data, which creates new requirements for methods of the processing and analysis. In this context, the concept of big data is used, which encompasses technologies for storing, processing and analytical modelling of very large amounts of information (Adamov, 2018). The main characteristics of modern information platforms are horizontal scalability, fault tolerance, high-speed streaming data processing and support for multi-user access. Structurally, modern corporate systems are increasingly built on the basis of cloud-oriented infrastructure, which ensures elastic use of computing resources and optimisation of the cost of operating the technological platform. A priority in the development of information systems is the integration of artificial intelligence (AI) methods, which transforms traditional information platforms into cognitive-oriented knowledge processing environments. Intelligent data analysis is used to identify hidden patterns, predict user behavioural characteristics, automate classification of information objects and support management decision-making. In corporate and web-based systems, machine learning algorithms can be incorporated as separate analytical services or as embedded business logic components, depending on the chosen architectural strategy (Emadi, 2023). The conceptual structure of a modern information platform, which demonstrates the relationship between the functional, data-centric, and intelligent components of the system, is given in Table 1.

Table 1. Conceptual levels of a modern information system

System level	Main purpose	Key technological elements
Infrastructure level	Provision of computing resources	Cloud services, containerisation, virtualisation
Service level	Implementation of business functions	API gateways, microservices, message queues
Data-driven level	Information storage and processing	Data warehouses, streaming analytics
Intellectual and analytical level	Cognitive information processing	Machine learning models, predictive analytics
Application level	User interaction	Web interfaces, mobile clients

Source: compiled by the author based on J. Emadi (2023), A. Efe (2024)

Table 1 reflects a stratified model of an information system, in which each level performs an autonomous but functionally related role in forming a holistic digital platform. This structuring corresponds to the principle of division of responsibilities and reduces architectural complexity, ensuring manageability, scalability and technological flexibility of the system. Conceptually, it is important that the intellectual and analytical component is presented as a separate level, which indicates the transition from transactional systems to cognitively oriented decision support platforms. Taken together, the model demonstrates the logic of sequential transformation of resources into applied value, forming the basis for building adaptive and scalable corporate ecosystems. Web-oriented information systems are characterised by a high level of client accessibility and platform independence. The use of multi-component architecture allows for the interaction of user interfaces with server logic through standardised programme interfaces. In this context, web platforms act as universal environments for integrating corporate resources, analytical services and external information sources. The use of intelligent algorithms in the web environment creates the prerequisites for the formation of adaptive interfaces, personalised recommendation mechanisms and automated monitoring of system behaviour. Thus, the conceptual principles of building modern information systems are based on a synergistic combination of architectural modularity, data-centric information flow management and artificial intelligence technologies, which ensures the transition from traditional automated platforms to intelligent digital ecosystems. A promising direction of development is the further formalisation of methods for integrating analytical models into the software-architectural contour of the system in order to increase its adaptability and operational efficiency.

Architectural models for designing modern information systems

Corporate information systems are designed as complex multi-tiered platforms, the architecture of which is aimed at ensuring functional isolation of components, high availability of services and the possibility of independent scaling of individual modules. The traditional client-server model evolved towards multi-tiered structures that involve the separation of presentation, business-logical and data-centric subsystems. Such decomposition allows reducing inter-component dependence, increasing the level of data processing security and ensuring flexibility in software modernisation without the need for complete system reconstruction. The transition from monolithic software solutions to service-oriented architectural models is associated with the development of the concept of Service-Oriented Architecture, which involves building a system as a set of loosely coupled services with standardised interaction protocols. Within this paradigm, the functional modules of the system encapsulate business logic and expose it through universal communication interfaces. A further development of this architectural idea was the Microservices architecture model, characterised by maximum autonomy of software components, the possibility of an independent service life cycle, and high horizontal scalability of the platform (Bondarchuk *et al.*, 2024).

A multi-tiered corporate system usually includes infrastructure, service, logical, and presentation layers. The infrastructure layer provides a physical or virtual environment for executing programme code, including storage resources, network communication, and computing power. The service layer implements the functionality of business processes through a set of interacting services, while the data layer is responsible for the consistency of information resources and transaction management. The presentation layer provides user interaction with the system through client interfaces. Comparative characteristics of architectural models of corporate systems are presented in Table 2.

The comparative matrix reflects the evolution of architectures from centralised to decentralised models, in which complexity is managed through decomposition and autonomisation of components. It demonstrates the transition from a tightly integrated structure to network interaction of independent services, which increases the flexibility and adaptability of the system. At the same time, the complexity does not disappear, but moves to the plane of coordination of distributed processes, data consistency management and infrastructure monitoring. Therefore, the choice of architecture is determined by the balance between scalability and organisational capacity to ensure effective management of the distributed environment. Conceptually, the interaction of corporate

platform components can be represented as a hierarchical structure of data flows and service calls. Information requests from client applications arrive at API gateways, after which these requests are routed to the corresponding business services, which process the data and form the result of the operation (Fig. 1).

Table 2. Comparative characteristics of architectural models of corporate systems

Criterion	Monolithic architecture	Service-oriented architecture	Microservice architecture
Scalability	Limited	Moderate	High
Modularity	Low	Medium	High
Difficulty of support	High as the system grows	Medium	Optimised
Component independence	Minimum	Partial	Maximum
Fault tolerance	Lower	Medium	High

Source: compiled by the authors based on J. Essien (2023), S.K. Jangam (2023)

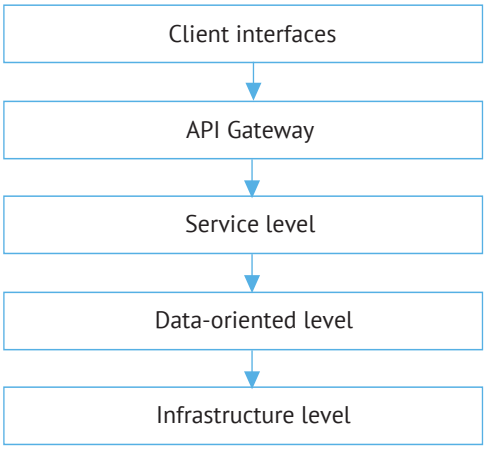


Figure 1. Hierarchical architectural model of query processing in a multi-level information system
Source: compiled by the authors based on H. Kozub & Yu. Kozub (2022), S.K. Jangam (2023), A. Matani *et al.* (2024)

The presented scheme reflects the logic of sequential transformation of information queries within a multi-tier architecture, where each tier acts as a functional intermediary between user interaction and underlying computing resources. Such a structural organisation provides centralised management of access to services and allows separating business operation processing mechanisms from infrastructure execution mechanisms. Conceptually, this model emphasises the principles of sequential abstraction and decomposition of the system, which contributes to the manageability of complex software platforms. It also creates the prerequisites for independent modernisation of individual layers of the system and integration of additional functional modules, in particular analytical or intelligent components, without violating the overall architectural integrity. Integration of corporate information resources is carried out through the use of standardised data exchange protocols, which allows for interaction between the organisation's internal subsystems and external information services. The use of event-driven interaction of components becomes important, which minimises synchronous system blocking and increases its performance in high-load environments.

Web-oriented information systems form a separate class of software platforms based on the principle of platform-independent user interaction with server logic via network data transfer protocols. The key architectural characteristic of such systems is a clear decomposition of frontend and backend components, which ensures the independence of user interface development from internal business logic. The frontend subsystem is responsible for implementing the presentation layer, forming dynamic interfaces, and ensuring interactive user interaction with the system. For this purpose, modern client frameworks are used that support a component-oriented development model and reactive interface state updating. The backend layer implements business process processing algorithms, database management, and integration with external services (Tekane, 2025).

Interaction between the client and server parts of the system is carried out mainly through API-oriented communication mechanisms based on the use of standard data exchange interfaces. The most common technological approach is the use of Representational State Transfer, which provides a unified model for accessing system resources via HTTP. The REST architecture involves the use of stateless requests, resource identification via URIs, and standardised operation methods. Cloud infrastructure plays a key role in ensuring the scalability of web-oriented platforms. The use of the Cloud computing concept allows for elastic scaling of computing resources, automatic load balancing, and

data redundancy (Liu *et al.*, 2023). Modern systems often use containerisation of software components, which provides a unified execution environment regardless of the physical infrastructure.

The integration of data intelligence modules into the overall architecture of the information system is based on the principle of multi-level analytical processing of information flows. Analytical components can be implemented as standalone services or as embedded algorithmic business logic modules, depending on the performance, security, and complexity requirements of machine learning models. Streaming and batch data processing form two fundamental paradigms of the analytical computing model. Batch processing is used to analyse large historical data sets and train predictive models, while streaming processing is focused on real time and involves incremental updating of analysis results. In this context, the concept of big data is actively used, which provides a technological basis for working with high-dimensional information flows. Integration of Machine learning algorithms into application services is carried out through the creation of specialised analytical microservices or through embedding predictive models directly into the business logic of the system (Muvva, 2025). This approach allows for the implementation of adaptive process optimisation functions, automated information classification, and detection of anomalous user behaviour patterns. A typical scheme for integrating analytical components can be represented as a sequence of technological levels (Fig. 2).

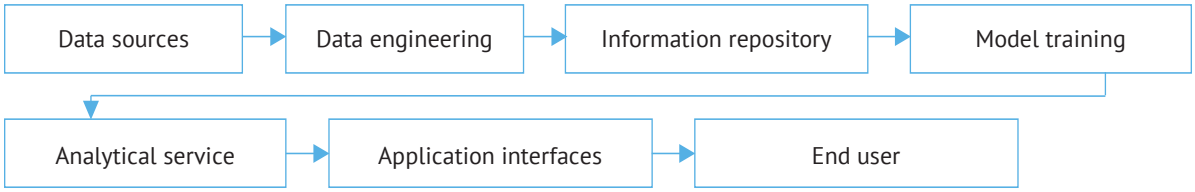


Figure 2. Integration scheme of analytical components
Source: compiled by the authors based on S. Muvva (2025)

The above structure demonstrates the shift of the architectural logic of information systems towards close integration of operational and analytical processes. Analytical components are built into the general information processing cycle, turning data not only into a storage resource, but into an active tool for supporting the functioning of the system. As a result, analytics becomes a component of application services and directly affects the formation of management or functional decisions. In such an architecture, data is the central element that combines infrastructure processing mechanisms with the applied logic of the system. This creates the prerequisites for the formation of adaptive information platforms that are able not only to respond to user requests, but also to predict behavioural patterns, optimise business processes and increase the efficiency of the entire digital ecosystem. Thus, the architectural integration of intelligent analytics transforms the information system into a cognitive-oriented platform capable of supporting automated decision-making in a dynamic digital environment. Further development of such systems is associated with the improvement of distributed model learning methods and optimisation of computational costs when operating intelligent services.

Methodological approaches to information systems design

The project-oriented methodology for designing information systems is based on the concept of a managed software product life cycle, which involves formalised management of all stages of creation, implementation, and operation of a digital platform. In modern software engineering, the system life cycle is considered as a continuous process that includes the stages of requirements analysis, architectural design, implementation of functional components, testing, deployment and further modernisation of the system. This approach provides the ability to adapt the information platform to changes in the technological environment and business requirements of the organisation. A key element of the project-oriented paradigm is the use of iterative software development models, which involve the gradual refinement of the functional specification of the system through a series of consecutive design and testing cycles. Among modern methodological approaches, flexible development methods based on the principles of adaptive project management are widespread. In this context, the Agile software development methodology ensures a high speed of response of the development team to changing customer requirements and promotes continuous improvement of the software product through short iterative cycles (Kundisch *et al.*, 2022).

Information system life cycle management can be formalised through a multi-level model of quality control and functional compliance of the system with user requirements (Table 3). In the structural dimension, the life cycle includes strategic architecture planning, technical implementation, operational monitoring and post-operational support of the software platform. The use of the principles of continuous integration and automated testing of software code is becoming increasingly important, which minimises the risks of critical system errors.

Table 3. Generalised structure of the information system life cycle

Life cycle stage	Main tasks	Resulting artifacts
Requirements analysis	Formalisation of functional needs	Requirements specification
Architectural design	Defining the structural model of the system	Architectural design
Implementation	Software development of components	Executable code
Testing	Function verification	Testing reports
Deployment	System deployment	Working platform
Support	Optimisation and modernisation	System update

Source: compiled by the authors based on D. Kundisch *et al.* (2022), N. Renaldo *et al.* (2022)

The presented structure reflects the system logic of organising the process of creating an information system, within which each stage forms the basis for the next level of technological implementation. It demonstrates the sequential transformation of the initial conceptual requirements into a functional software product through a system of interconnected engineering procedures. This approach emphasises that the effectiveness of development is determined not only by technical solutions, but also by the quality of management of the design, control, and improvement processes of the system. In general, the table indicates a transition from a linear understanding of the life cycle to a cyclical and adaptive model of software platform development. This provides the possibility of continuous improvement of the system architecture, integration of new functional components and maintenance of stability of work in a changing technological environment. Iterative software development methodologies involve the use of incremental increase in system functionality, which makes it possible to reduce the complexity of managing large projects. The development process is divided into a series of short cycles, within which planning, implementation, evaluation of results and correction of architectural solutions are carried out. This model minimises technological risks and ensures more stable integration of software modules into complex information environments.

Domain-driven design of information systems is based on the principle of structural correspondence of the software architecture to the subject area of the system's functioning. The main idea of this approach is the maximum coordination of the business logic of the software product with the conceptual model of the organisational environment in which the system is operated. Within the framework of this methodology, the software system is considered as a formalised model of business processes that reflects the structural and functional characteristics of the subject area. The concept of Domain-driven design involves the hierarchical organisation of the software code in accordance with the business domains of the organisation (Zhang, 2025). The architectural structure of the system is formed around key business objects that determine the logic of information processing and interaction between functional modules (Hackenbeck *et al.*, 2025). Within the framework of this methodology, the concepts of domain core, aggregates, domain services and data repositories are distinguished. Aligning the software architecture with the subject area reduces the cognitive complexity of the software and increases the level of scalability of the system. An important aspect is the formation of a universal business logic model that can be reused when upgrading the information platform or integrating new functional components. The interaction of the elements of a domain-oriented architecture can be represented in the form of a diagram (Fig. 3).

Figure 3. Interaction of elements of domain-oriented architecture

Source: compiled by the authors based on C. Hackenbeck *et al.* (2025)

The presented scheme reflects the conceptual translation of real business processes into the software architecture of the system, where the domain model plays a central role as an intellectual framework of applied logic. Such a structural organisation demonstrates that the software system is formed not around technical components, but around the substantive elements of the subject area, which provides a more natural reflection of organisational processes in the software environment. In general, this model emphasises the importance of semantic consistency between the conceptual description of the organisation's activities and the software implementation of the information platform. This creates the prerequisites for increasing the clarity of the system architecture, simplifies its evolutionary development, and ensures the stability of business logic when expanding or modifying functional components.

The data-driven architectural strategy for designing information systems involves the use of analytical methods of information processing to form optimal parameters of the software platform. This approach is based on the concept of decision-making based on empirical data, which allows minimising the subjective factor when choosing technological solutions. Within

the framework of this paradigm, data mining methods are actively used, which ensure the detection of hidden patterns in large amounts of information. Analytical results can be used to optimise the structure of databases, balance the load between service nodes, and predict the behavioural characteristics of system users (Hensel *et al.*, 2022). Using analytics to optimise the functional components of the information platform involves constant monitoring of the system's operational metrics, including bandwidth, query processing latency, service error rate, and the efficiency of computing resources. Based on the collected telemetric data, automated correction of the configuration parameters of the software environment is carried out, which forms the basis of the concept of self-tuning information systems. A typical structure of data-driven architectural management can be represented as a sequence of analytical operations (Fig. 4).

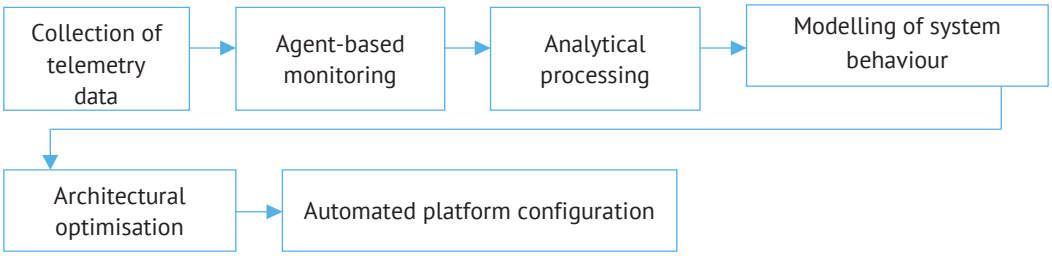


Figure 4. Typical structure of data-driven architectural management
Source: compiled by the author based on K. Kayabay *et al.* (2022)

The presented scheme reflects the transition to an architectural model in which the design and optimisation of the information system are carried out on the basis of continuous analysis of operational data. In such logic, the results of monitoring the functioning of the platform become the basis for correcting configuration parameters, improving the structural organisation of components and increasing the efficiency of resource use. In general, the model demonstrates the formation of an adaptive information infrastructure capable of self-regulation and evolutionary improvement in accordance with changes in load, user behaviour and technological conditions of the system's operation. Thus, methodological approaches to the design of information systems form an integrated system of engineering and analytical management of the life cycle of software platforms, combining the principles of iterative development, domain decomposition and empirically based architectural decision-making. Further development of design methodologies is associated with the implementation of autonomous intelligent mechanisms for optimising software systems.

Practical aspects of implementing information systems with intelligent analytics

The practical implementation of information systems with integrated analytical modules involves the formation of a comprehensive technological platform that combines infrastructure, service and cognitive components of information processing. In modern conditions of digital transformation, corporate information systems are considered as multifunctional knowledge management environments that can provide not only the automation of operational processes, but also support intelligent decision-making based on the analysis of large volumes of data. The basis for building such platforms is the concept of integrated architecture, which ensures the interaction of data sources, analytical algorithms and application services through standardised software interfaces. The conceptual model of a corporate information platform can be represented as a multi-level structure that includes the data collection level, the level of the storage and pre-processing, the analytical level of knowledge modelling and the level of applied interpretation of results. At the infrastructure level, information flows from internal corporate systems, external web sources, and sensor or transactional subsystems are accumulated. For effective management of large amounts of information, the data warehouse concept is used, which provides structured storage of historical data and support for highly complex analytical queries. Distributed information processing technologies can be used in parallel, in particular ecosystems based on the principles of big data (Ismail *et al.*, 2024).

The service level of the platform includes analytical modules that implement algorithmic processing of information flows and the formation of predictive models of system behaviour. In this context, Machine learning methods are used to automate the processes of detecting patterns in high-dimensional data. The practical architecture of such platforms is often based on the concept of microservice interaction of analytical components, where each algorithmic module functions as an autonomous knowledge processing service. An example of a web-based information system with an analytical module can be implemented as an integrated user data management platform that provides personalisation of services and prediction of user behavioural characteristics. Such systems use the technology of intelligent analysis of web activity, which allows forming dynamic user profiles based on the history of interaction with

the digital environment. The interface part of the platform implements mechanisms for visualising analytical results, which helps to increase the interpretability of machine learning models and support management decision-making (Hossain *et al.*, 2024). A typical functional structure of an information system with intelligent analytics can be formalised in the form of a multilayer architecture of component interaction (Table 4).

Table 4. Functional levels of the analytical information platform

Platform level	Functional purpose	Key technological solutions
Data sources	Accumulation of information flows	API integration, streaming brokers
Data processing	Preliminary information transformation	Extract, Transform, Load/Extract, Load, Transform processes
Analytical level	Intelligent modelling	ML models, statistical analytics
Service level	Application business logic	Microservices, RESTful services
Presentation of results	Analytics visualisation	Dashboard systems

Source: compiled by the authors based on Q. Hossain *et al.* (2024)

The table below summarises the logic of the functioning of analytically oriented information platforms, demonstrating the transition from traditional transactional systems to architectures in which analytical data processing is integrated directly into operational processes. In such a model, information flows undergo a sequential transformation – from primary accumulation to intelligent interpretation and applied use of results, which forms a single data management loop. The synthesis of the presented levels indicates the formation of a holistic data-centric architecture in which analytical algorithms become an integral element of the platform's service logic, ensuring automation of decision-making, personalisation of digital services and increased adaptability of corporate information systems to changes in the information environment.

The implementation of intelligent data analysis functions involves the use of different types of algorithmic models depending on the information processing task. Forecasting behavioural characteristics is carried out by building temporal or stochastic models that allow extrapolating future values of the studied parameters based on historical observations. In this context, predictive analytics methods are used, based on the mathematical apparatus of statistical learning and neuroevolutionary modelling. The use of time series analysis algorithms for processing dynamic information flows is important. The tasks of classification and segmentation of information data are solved through the use of structural learning methods, which provide grouping of information objects according to the degree of similarity of the characteristics. For this, clustering, discriminant analysis and deep learning algorithms can be used, which allows the formation of multidimensional data distribution models. Personalised recommendation mechanisms are based on the analysis of user behaviour history and involve the use of the principles of collaborative filtering or content-oriented interaction modelling (Arafat *et al.*, 2025).

Detection of anomalies in information systems is a critically important functional task, as it allows ensuring cybersecurity and stable functioning of digital platforms. In this aspect, statistical process control methods and machine learning algorithms are used to identify deviations from the normal operating mode of the system. The practical implementation of such mechanisms may include the use of Anomaly detection models, which are capable of automatically recording atypical patterns of user behaviour or system processes (Usmani *et al.*, 2024). Thus, the practical implementation of information systems with intelligent analytics forms the technological basis for creating adaptive digital platforms of a new generation that combine the capabilities of distributed data processing, predictive modelling and automated support for making management decisions. The further development of such systems will be determined by the improvement of model training algorithms and increasing the interpretability of intelligent services.

Evaluating the effectiveness of architectural solutions for information systems

Evaluating the effectiveness of architectural solutions for information systems is a complex multidimensional task that involves analysing the technological, functional, and analytical characteristics of the software platform. In modern software engineering, system efficiency is considered an integral property of the architecture, which reflects the ability of the information platform to provide stable data processing at minimal computational costs and a high level of accuracy of analytical results. Methodologically, the effectiveness assessment is based on the principles of quantitative performance metrics, statistical validation of models, and analysis of system reliability under variable load conditions.

Information systems performance metrics form the basis of the technical assessment of architectural solutions and allow determining the optimal parameters of the software platform's functioning. Key performance indicators include request processing time, system throughput, network interaction latency, and the efficiency of computing resources. In the context of distributed platforms, the concept of Cloud computing performance metrics is of particular importance, which is used to assess the scalability of cloud services, the level of resource availability and the variability of the system's response time when the number of user requests increases. The scalability of an information system is determined by the ability of the architecture to support

the growth of the computing load without critical deterioration in performance. Modern digital platforms use the principle of horizontal scaling, which involves adding new computing nodes to a distributed cluster environment. In this context, Distributed computing technologies are actively used, which provide parallel processing of information flows and load balancing between service components of the system (Hackenbeck *et al.*, 2025). The main performance metrics of information systems are presented in Table 5.

Table 5. Key metrics for information systems performance

Metrics group	Indicator	Measurement description
Time characteristics	Latency	Average system response time
	Throughput	Number of requests processed per unit of time
Resource characteristics	Central processing unit utilisation	Central processing unit resource loading
	Memory consumption	Random access memory usage
Reliability	Error rate	Frequency of system errors
Scalability	Scalability coefficient	Performance dynamics with increasing load

Source: compiled by the authors based on M.S. Sultana & S. Akter (2023), P.R. Nangi *et al.* (2024)

The presented Table 5 systematises approaches to quantitative assessment of the functioning of information platforms, demonstrating that the effectiveness of architectural solutions is formed not by separate technical indicators, but by the coordinated set. The combination of time, resource, and reliability characteristics reflects the multidimensional nature of the performance of distributed systems, in which the stability of query processing, the efficiency of the use of computing resources and resistance to load growth are interconnected. In this context, metrics serve as tools for empirical verification of architectural solutions, allowing to assess how much the selected technological model provides a balance between speed, economy of infrastructure use and reliability of system operation in conditions of scaling and dynamic load changes.

Assessing the accuracy of analytical models is a critically important aspect of the functioning of information systems with integrated intelligent components. Verification of predictive and classification algorithms is carried out by comparing model results with empirical observations. In this context, machine learning methods based on the mathematical apparatus of statistical inference are used. A significant role is played by the use of Supervised learning approaches, which involve training models based on pre-labelled data (Kumar *et al.*, 2022). Standard forecast error metrics are used to quantitatively assess the quality of analytical models (Table 6). Among these forecast error metrics, the most common are the mean absolute error, the mean square error, and the relative forecast error. These indicators allow formalising the level of modelling accuracy and providing an objective comparison of alternative algorithmic solutions.

Table 6. Metrics for assessing the accuracy of analytical models

Metrics	Mathematical content	Scope of application
MAE	Mean absolute deviation	Robust forecast estimate
RMSE	Root-mean-square forecast error	Sensitivity to large deviations
MAPE	Relative forecast error	Normalised accuracy estimate
Precision	True positive classification rate	Binary classification
Recall	Object detection completeness	Detection tasks

Notes: MAE – mean absolute error; RMSE – root mean squared error; MAPE – mean absolute percentage error
Source: compiled by the authors based on R.C. Khamari *et al.* (2025)

The presented system of metrics reflects the desire to formalise the assessment of the quality of intelligent models by combining statistical and classification criteria. The use of different types of errors allows for multidimensional verification of predictive algorithms, since individual indicators are sensitive to different characteristics of data distribution and the structure of modelling errors. In general, such a composition of metrics forms a universal toolkit for validating analytical systems, which can be used both for forecasting tasks and for tasks of structural classification of information objects, ensuring a balance between the model's resistance to outliers, prediction accuracy and completeness of detection of relevant patterns. The analysis of the reliability of an information system involves studying the stability of the software platform's functioning in conditions of component failures or external destructive influences. The reliability of the architecture is determined by the probability of failure-free operation of the system during a given time interval. In modern digital environments, considerable attention is paid to the concept of Fault-tolerant system, which involves the implementation of mechanisms for redundancy, automatic restoration of services and isolation of erroneous processes. The adaptability of the information platform characterises the ability of the system to change the parameters of its functioning in accordance with variations in the external environment or user behaviour. In this

aspect, the principles of self-tuning computing environments are applied, which can automatically optimise the configuration of the software infrastructure based on telemetric data. The overall integrated efficiency of the information system architecture can be assessed through a combined performance index that aggregates the time, resource and analytical characteristics of the platform. The formation of such an index allows for a comparative analysis of alternative architectural models and to justify the choice of a technological strategy for the development of the information system.

DISCUSSION

The analysis of scientific publications indicates the active development of the concept of integrating analytics into the architecture of information systems, which is consistent with the results of the study on the formation of data-centric adaptive digital platforms. Thus, in the work of M.C. Solano & J.C. Cruz (2024), a systematic review of the integration of analytical components into corporate information systems was carried out, where it was shown that the implementation of analytics increases operational efficiency, supports decision-making and promotes the digital transformation of enterprises. The conclusions obtained by the authors correlate with the results of the study on the feasibility of architectural integration of intelligent data analysis into the service contour of the platform. At the same time, the specified work details the structural organisation of multi-level architectural models to a lesser extent, which explains the differences in the level of conceptual formalisation. The study of A. Almoqry & N. ALSohybe (2025) considered the architectural aspects of the digital transformation of information systems in the field of education, where the importance of integrated data management platforms and flexible architectural models was highlighted. The authors' results confirmed the prospects of using adaptive information environments, but the emphasis is mainly on the organisational and institutional conditions for the implementation of digital technologies. Compared to the current study, which focuses on theoretical modelling of the architectural logic of the system and the integration of analytical algorithms into the technological contour of the platform, the work of A. Almoqry & N. ALSohybe demonstrates a more applied and managerial focus, which causes a partial difference in the results. In turn, the publication of V.-H. Capacho-Alfonso *et al.* (2025) is devoted to a bibliometric analysis of the development of enterprise architecture in the context of smart cities, healthcare and new technologies. The authors established the growth of research related to the use of artificial intelligence and architectural innovations in digital ecosystems. The obtained trends are generally consistent with the current conclusion about the dominance of cognitively oriented data processing platforms. At the same time, the differences are that the bibliometric approach captures only structural trends in the development of the industry, while the conducted research is conceptual and methodological in nature and is aimed at forming a model of architectural integration of analytical components.

The impact of generative artificial intelligence on the activities of information systems architects in Agile environments is studied in the work of S.J. Kooy *et al.* (2025). The authors showed that the use of generative models can support design processes, but emphasise the need for quality control of generated architectural solutions and human expert verification support. The conclusions obtained are consistent with the results of the conducted research on the feasibility of integrating intelligent analytics into the technological contour of information platforms. The study by J. Weller & R. Dumitrescu (2025) proposed a decision-driven analytics approach for designing information systems in a smart manufacturing environment. The authors pointed out the importance of integrating analytical models directly into the architecture of business processes, which correlates with the current conclusion about the formation of data-centric platforms, in which analytics acts as a functional component of the operational loop. The differences between the results lie in the level of detail: the aforementioned study is focused on the engineering of use-case scenarios for industrial systems, while the analysis focuses on a generalised theoretical and methodological description of the architectural logic of information ecosystems. At the same time, the publication by N.R. Busch & A. Zalewski (2025) is devoted to a review of methods for evaluating enterprise architecture. The authors noted that the effectiveness of architectural models is determined by the balance between system complexity, scalability, reliability, and manageability of the software life cycle. These results directly confirmed the conclusions obtained regarding the multidimensional assessment of the performance of information platforms.

Interrelated aspects of the development of data-driven digital environments, in particular the technological optimisation of big data processing and the organisational and architectural impact of analytics on the innovative potential of information systems were demonstrated by the works of V. Winata (2022) and S. Pathak *et al.* (2025). Taken together, these studies confirm the general scientific position on the growing importance of artificial intelligence in managing information flows and supporting the adaptive functioning of digital platforms. At the same time, the conducted study expands this direction by conceptually and methodologically generalising the principles of integrating intelligent analytics into the architecture of information systems, which makes it possible to consider analytical modules as a structural component of a digital platform, and not only as a tool for technological optimisation. The study by B. Adhytia *et al.* (2022) demonstrated that enterprise architecture is a key factor in realising the value of big data analytics in organisational systems. The authors emphasised that the effectiveness of analytical platforms depends on the consistency of the infrastructure, service, and information layers of the architecture.

These findings support the results obtained regarding the multi-level stratified model of the information system. At the same time, the current work extends existing approaches by including the intellectual-analytical layer as a separate functional layer of the architecture, which is less detailed in the aforementioned studies. The study by A. Erica *et al.* (2024) shows that the use of data analytics in management systems contributes to improving the quality of management decision-making through structured interpretation of information flows. The authors stressed the role of integrated analytics in supporting management-oriented information systems, which is consistent with the current conclusion about the need to embed analytical modules into the operational contour of the platform.

The use of artificial intelligence to create adaptive learning environments in digital educational platforms is investigated in the publication by I. Gligorea *et al.* (2023). The authors proved that adaptive learning mechanisms increase the personalisation of digital interaction and improve the efficiency of information services. The results obtained partially correlate with the current conclusion about the adaptability of information systems as a key characteristic of modern digital platforms. The difference lies in the industry specificity: the conducted study is of a universal nature and is aimed at generalising the architectural principles of corporate and web-oriented information systems, while the mentioned work focuses on educational systems. In turn, the work of H.C. Ling & H.S. Chiang (2022) investigated the effectiveness of adaptive recommendation systems in the educational web environment. The authors showed that personalised recommendation algorithms have a positive effect on user productivity indicators by dynamically taking into account behavioural patterns of interaction with the system. The results obtained are consistent with the obtained conceptual conclusion about the importance of adaptability of information platforms.

In the publication of A. Gaurav *et al.* (2023), the authors considered the application of machine learning to detect malicious software in the infrastructures of the Internet of Things of corporate information systems. The authors confirmed the effectiveness of intelligent algorithms in the tasks of cyber protection of digital platforms, which correlates with the current theoretical position on the importance of integrating analytical mechanisms to ensure the stability of information systems. At the same time, the study by A. Gaurav *et al.* (2023) has a highly specialised technical focus, while the conducted study considered more general architectural and methodological principles of applying intelligent analytics in information platforms.

A comparison of the results obtained showed that most studies confirm the importance of integrating analytical and intelligent components into the architecture of information systems, in particular in the context of adaptability of digital platforms, personalised data processing and the use of machine learning methods to improve functional efficiency. A general trend of transition from local algorithmic solutions to a systemic architectural vision of information ecosystems, which corresponds to the concept of data-centric and cognitive-oriented platforms, was identified. The differences between the results of other authors and the presented study are mainly due to the level of generalisation – in the analysed works, the emphasis was on empirical experiments or highly specialised applied problems, while in this study, a conceptual and methodological synthesis of architectural approaches to the design of information systems was carried out. Thus, the results obtained are consistent with modern scientific trends in the development of digital platforms, but expand these trends due to the integrated theoretical substantiation of the principles of designing adaptive information systems.

CONCLUSIONS

The analysis found that modern information platforms are developing in the direction of transition from monolithic architectures to multi-level modular models that provide scalability, functional isolation of components and technological flexibility of systems. Information systems are increasingly integrating analytical components into the overall architectural contour, transforming from automation tools into intelligent decision-making support environments. The use of big data and artificial intelligence algorithms ensures the processing of complex information flows, the identification of patterns and the prediction of user behaviour. The design of information platforms is based on an iterative software development life cycle, which includes requirements analysis, architectural modelling, implementation, testing, and modernisation of the system. Domain-oriented design promotes semantic consistency of the software architecture with the subject area of the system, reduces software complexity and increases scalability. The data-driven approach makes it possible to use operational and telemetric data to optimise the platform configuration and load balancing between services.

The practical implementation of intelligent information systems involves a multi-level data processing structure that covers the collection, storage, analytical processing, modelling, and visualisation of information. The effectiveness of the architecture is assessed comprehensively using time, resource, and reliability characteristics. Scalability is ensured by horizontal expansion of the computing infrastructure and the use of the metrics latency, throughput, central processing unit utilisation, memory consumption, error rate and scalability factor. The quality of analytical models is assessed using the metrics MAE, RMSE, MAPE, precision and recall. The reliability of information systems is ensured by fault-tolerant architectural solutions, service redundancy, automatic process recovery and the use of self-tuning computing environments. The results obtained confirm the prospects for the development of adaptive cognitive-oriented information systems that combine architectural modularity, intelligent analytics and effective management of digital resources. The

limitations of the study are its predominantly theoretical and analytical nature without conducting empirical experimental evaluation of the proposed models in real information environments. Prospects for further research are related to the practical validation of conceptual approaches by modelling architectural solutions of information systems and assessing the effectiveness under real load conditions.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Abdulrahman, L., Zeebaree, S.R.M., & Omar, N. (2022). State of art survey for designing and implementing regional tourism web based systems. *Academic Journal of Nawroz University*, 11(3), 100-112. doi: 10.25007/ajnu.v11n3a1425.

[2] Adamov, A. (2018). Large-scale data modelling in Hive and distributed query processing using Mapreduce and Tez. In *DiVAI 2018. 12th international scientific conference on distance learning in applied informatics* (pp. 389-404). Sturovo: Wolters Kluwer.

[3] Adhytia, B., Kurnia, S., Dilnutt, R., & Hidayanto, A.N. (2022). Enterprise architecture contribution to big-data analytics value realization: A literature analysis. *PACIS 2022 Proceedings*, article number 189.

[4] Almoqry, A., & ALSohybe, N. (2025). Enterprise architecture in higher education for digital transformation: An integrative review for fragile contexts with application to Yemen. *F1000Research*, 14, article number 1192. doi: 10.12688/f1000research.172049.1.

[5] Arafat, Y., Akula, D.K., Mohammed, Y.S., Haque, G.M., Rahman, M.B., & Syed, A. (2025). Big data analytics in information systems research: Current landscape and future prospects focus: Data science, cloud platforms, real-time analytics in IS. *Emerging Frontiers Library for The American Journal of Engineering and Technology*, 7(8), 177-201. doi: 10.37547/tajet/Volume07Issue08-16.

[6] Bondarchuk, O., Kozub, V., & Kozub, Yu. (2024). Analysis of the effectiveness of machine learning algorithms in big data processing. *Information Technology and Society*, 4(15), 71-79. doi: 10.36910/6775-2524-0560-2024-56-13.

[7] Busch, N.R., & Zalewski, A. (2025). A systematic literature review of enterprise architecture evaluation methods. *ACM Computing Surveys*, 57(5), article number 113. doi: 10.1145/3706582.

[8] Capacho-Alfonso, V.-H., Jiménez-Builes, J.A., & Duran, D.E. (2025). Aspects of enterprise architecture in emerging technologies, smart cities and the health sector: A bibliometric study. *Salud, Ciencia y Tecnología*, 5, article number 1656. doi: 10.56294/saludcyt20251656.

[9] Efe, A. (2024). Evolution of management information systems by super artificial intelligence revolutions. *International Journal of Management Information Systems and Computer Science*, 8(2), 127-142. doi: 10.33461/uybisbbd.1521086.

[10] Emadi, J. (2023). The development of a design theory for web based information systems. *Journal of Robotics Spectrum*, 1, 13-23. doi: 10.53759/9852/JRS202301002.

[11] Erica, A., Gantari, L., Qurotulain, O., Nuche, A., & Sy, O. (2024). Optimizing decision-making: Data analytics applications in management information systems. *APTISI Transactions on Management*, 8(2), 115-122. doi: 10.33050/atm.v8i2.2202.

[12] Essien, J. (2023). Enterprise architecture: A comparative analysis of validation semantics and heterogeneous model frameworks. *Open Journal of Business and Management*, 11(5), 1971-1995. doi: 10.4236/ojbm.2023.115109.

[13] Gaurav, A., Gupta, B.B., & Panigrahi, P.K. (2023). A comprehensive survey on machine learning approaches for malware detection in IoT-based enterprise information system. *Enterprise Information Systems*, 17(3), article number 2023764. doi: 10.1080/17517575.2021.2023764.

[14] Gligorea, I., Cioca, M., Oancea, R., Gorski, A.T., Gorski, H., & Tudorache, P. (2023). Adaptive learning using artificial intelligence in e-learning: A literature review. *Education Sciences*, 13(12), article number 1216. doi: 10.3390/educsci13121216.

[15] Hackenbeck, C., Walz, A., Hirmer, P., Wagner, S., & Weyrich, M. (2025). Semantic interface modeling for automotive architectures using a domain-driven approach. *IFAC-PapersOnLine*, 59(25), 125-130. doi: 10.1016/j.ifacol.2025.11.936.

[16] Hensel, D.S., Tyc, J., & Hensel, M. (2022). Data-driven design for architecture and environment integration. *Spool*, 9(2), 19-34. doi: 10.47982/spool.2022.1.02.

[17] Hossain, Q., Yasmin, F., Biswas, T.R., & Asha, N.B. (2024). Integration of big data analytics in management information systems for business intelligence. *Saudi Journal of Business and Management Studies*, 9(9), 192-203. doi: 10.36348/sjbms.2024.v09i09.002.

[18] Ismail, A., Setiawati, R., Herbenita, H., Sutejo, B., & Mulyanto, S. (2024). *Management information system: Practical insights and applications in Indonesia*. Boulder: Asadel Publisher.

[19] Jangam, S.K. (2023). Data architecture models for enterprise applications and their implications for data integration and analytics. *International Journal of Emerging Trends in Computer Science and Information Technology*, 4(3), 91-100. doi: 10.63282/3050-9246.IJETCSIT-V4I3P110.

[20] Kayabay, K., Gökalp, M.O., Gökalp, E., Eren, P.E., & Koçyiğit, A. (2022). Data science roadmapping: An architectural framework for facilitating transformation towards a data-driven organization. *Technological Forecasting and Social Change*, 174, article number 121264. doi: 10.1016/j.techfore.2021.121264.
[21] Khamari, R.C., Mani, S., Bodkhe, R.G., & Singh, A.K. (2025). Enhancing wind power forecasting accuracy: A hybrid SNGF-RERNN-SCSO approach. *Solar Energy*, 295, article number 113513. doi: 10.1016/j.solener.2025.113513.
[22] Khudov, H., Khizhnyak, I., Salnyk, O., Mynko, P., & Andronov, V. (2025). Analysis of approaches to modeling and architectural representation of image processing information technology from onboard opto-electronic observation systems. *Control, Navigation and Communication Systems*, 3(81), 151-158. doi: 10.26906/SUNZ.2025.3.151.
[23] Kooy, S.J., Piest, J.P., & Bemthuis, R.H. (2025). Impact and implications of generative AI for enterprise architects in agile environments: A systematic literature review. *ArXiv*. doi: 10.48550/arXiv.2510.22003.
[24] Kozub, H., & Kozub, Yu. (2022). Declarative methodforcreating multiplatform applications. *Visnik of the Volodymyr Dahl East Ukrainian National University*, 5(275), 10-15. doi: 10.33216/1998-7927-2022-275-5-10-15.
[25] Kumar, S., Kolekar, T., Kotecha, K., Patil, S., & Bongale, A. (2022). Performance evaluation for tool wear prediction based on Bi-directional, Encoder-Decoder and Hybrid Long Short-Term Memory models. *International Journal of Quality & Reliability Management*, 39(7), 1551-1576. doi: 10.1108/IJQRM-08-2021-0291.
[26] Kundisch, D., Muntermann, J., Oberländer, A.M., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2022). An update for taxonomy designers: Methodological guidance from information systems research. *Business & Information Systems Engineering*, 64(4), 421-439. doi: 10.1007/s12599-021-00723-x.
[27] Ling, H.C., & Chiang, H.S. (2022). Learning performance in adaptive learning systems: A case study of web programming learning recommendations. *Frontiers in Psychology*, 13, article number 770637. doi: 10.3389/fpsyg.2022.770637.
[28] Liu, C.H., Chen, S.L., & Huang, H.K. (2023). Automated test input generation for testing representational state transfer (REST) application programming interface (API) using parameter fuzzing. In *Proceedings of the 6th international conference on knowledge innovation and invention* (pp. 249-253). Pekin: IEEE. doi: 10.1109/ickii58656.2023.10332662.
[29] Luo, J., Xu, J., Aldosari, O., Althubiti, S.A., & Deebani, W. (2022). Design and implementation of an efficient electronic bank management information system based data warehouse and data mining processing. *Information Processing & Management*, 59(6), article number 103086. doi: 10.1016/j.ipm.2022.103086.
[30] Matani, A., Sahafi, A., & Broumandnia, A. (2024). Improving query processing in blockchain systems by using a multi-level sharding mechanism. *The Journal of Supercomputing*, 80(10), 15066-15096. doi: 10.1007/s11227-024-06037-5.
[31] Molodets, B., & Bulana, T. (2024). Comparison of the capabilities of service-oriented architecture and microservice architecture in the creation of information systems. *Information Technologies in Metallurgy and Machine Building*, 17, 170-173. doi: 10.34185/1991-7848.itmm.2024.01.029.
[32] Muslih, M., Maulana, M.R., & Arianti, N.D. (2024). Development of a web-based job and career compatibility system using the federal enterprise architecture framework method: A case study in Nusa Putra university. *Ingénierie des Systèmes d'Information*, 29(3), 895-908. doi: 10.18280/isi.290310.
[33] Muvva, S. (2025). Bridging the gap: Integrating batch and streaming data paradigms for holistic analytics in the age of real-time, predictive, and historical insights. *International Journal of Communication Networks and Information Security*, 17(2), 330-358.
[34] Nangi, P.R., Obannagari, C.K., & Settipi, S. (2024). Serverless computing optimization strategies using ML-based auto-scaling and event-stream intelligence for low-latency enterprise workloads. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(3), 131-142. doi: 10.63282/3050-9246.IJETCSIT-V5I3P113.
[35] Obukhov, V., & Tovbych, V. (2025). Modelling of architectural design processes by means of artificial intelligence technologies. *Current Problems of Architecture and Urban Planning*, 71, 197-211. doi: 10.32347/2077-3455.2025.71.197-211.
[36] Pathak, S., Krishnaswamy, V., & Sharma, M. (2025). A dynamic capability perspective on the impact of big data analytics and enterprise architecture on innovation: An empirical study. *Journal of Enterprise Information Management*, 38(2), 532-563. doi: 10.1108/JEIM-01-2024-0059.
[37] Pinheiro, C.R., Guerreiro, S.L., & Mamede, H.S. (2024). A lightweight ontology for enterprise architecture mining of API gateway logs. *IEEE Access*, 12, 128585-128601. doi: 10.1109/ACCESS.2024.3456119.
[38] Rak, T., & Żyła, R. (2022). Using data mining techniques for detecting dependencies in the outcoming data of a web-based system. *Applied Sciences*, 12(12), article number 6115. doi: 10.3390/app12126115.
[39] Renaldo, N., Suhardjo, S., Suharti, S., Suyono, S., & Cecilia, C. (2022). Benefits and challenges of technology and information systems on performance. *Journal of Applied Business and Technology*, 3(3), 302-305. doi: 10.35145/jabt.v3i3.114.
[40] Solano, M.C., & Cruz, J.C. (2024). Integrating analytics in enterprise systems: A systematic literature review of impacts and innovations. *Administrative Sciences*, 14(7), article number 138. doi: 10.3390/admsci14070138.
[41] Sultana, M.S., & Akter, S. (2023). High-performance computing for scaling large-scale language and data models in enterprise applications. *ASRC Procedia: Global Perspectives in Science and Scholarship*, 3(1), 94-131. doi: 10.63125/e7yfwm87.

[42] Tekane, C.S. (2025). *A modular and configurable web-based frontend architecture for educational modelling tools*. (Master's thesis, McGill University, Montreal, Canada).
[43] Usmani, U.A., Aziz, I.A., Jaafar, J., & Watada, J. (2024). Deep learning for anomaly detection in time-series data: An analysis of techniques, review of applications, and guidelines for future research. *IEEE Access*, 12, 174564-174590. doi: 10.1109/ACCESS.2021.3107975.
[44] Vinueza-Martinez, J., Correa-Peralta, M., Ramirez-Anormaliza, R., Franco Arias, O., & Vera Paredes, D. (2024). Geographic information systems (GISs) based on WebGIS architecture: Bibliometric analysis of the current status and research trends. *Sustainability*, 16(15), article number 6439. doi: 10.3390/su16156439.
[45] Weller, J., & Dumitrescu, R. (2025). Decision-driven analytics in smart factories: Enterprise architecture framework for use case specification and engineering (FUSE). *Electronics*, 14(21), article number 4271. doi: 10.3390/electronics14214271.
[46] Winata, V. (2022). Optimizing big data processing through artificial intelligence: A systematic literature review. *Artificial Intelligence Research and Applied Learning*, 1(2). doi: 10.1234/aira.v1i2.36.
[47] Zhang, Y. (2025). Exploration of enterprise big data microservice architecture based on domain-driven design (DDD). *ArXiv*. doi: 10.48550/arXiv.2511.05880.

Архітектурні та методологічні підходи до проектування корпоративних і веборієнтованих інформаційних систем з елементами інтелектуального аналізу даних

Владислав Козуб

Доктор філософії з комп'ютерних наук, доцент
Луганський національний університет імені Тараса Шевченка
37500, вул. Віктора Новікова, 2, м. Лубни, Україна
https://orcid.org/0000-0003-2710-7206

Ольга Смагіна

Кандидат педагогічних наук, доцент
Луганський національний університет імені Тараса Шевченка
37500, вул. Віктора Новікова, 2, м. Лубни, Україна
https://orcid.org/0000-0002-6024-5152

Галина Козуб

Кандидат технічних наук, доцент
Луганський національний університет імені Тараса Шевченка
37500, вул. Віктора Новікова, 2, м. Лубни, Україна
https://orcid.org/0000-0001-5387-050X

Анотація. Актуальність дослідження зумовлена цифровою трансформацією та необхідністю удосконалення архітектурних і методологічних підходів до проєктування інтелектуальних корпоративних і веборієнтованих інформаційних систем. Метою дослідження було сформувати концептуальний підхід до проєктування інформаційних систем із застосуванням інструментів інтелектуального аналізу даних. Дослідження ґрунтувалося на теоретичному аналізі наукових матеріалів та застосуванні системного, концептуального, проєктно-орієнтованого, доменно-орієнтованого та даноорієнтованого методологічних підходів для обґрунтування принципів проєктування адаптивних інформаційних систем. Отримані результати засвідчили перехід сучасних інформаційних систем до багаторівневих модульних та даноцентричних архітектур, які поєднують функціональну гнучкість, масштабованість та інтелектуальну аналітичну обробку даних. Інтеграція алгоритмів штучного інтелекту забезпечує реалізацію прогнозної аналітики, класифікації інформації та виявлення аномалій у поведінці системи. Ефективність платформ оцінюється через комплекс часових, ресурсних та аналітичних метрик, зокрема затримкою, пропускною здатністю, рівнем використання центрального процесора, споживанням оперативної пам'яті, частотою помилок, середньою абсолютною похибкою, середньоквадратичною похибкою, середньою абсолютною відсотковою похибкою, точністю та повнотою. Показано, що надійність інформаційних систем забезпечується використанням відмовостійких архітектур, резервуванням сервісних компонентів та самонастроюваними обчислювальними середовищами. Узагальнення результатів підтвердило перспективність розвитку адаптивних когнітивно-орієнтованих даноцентричних інформаційних платформ нового покоління. Результати можуть бути використані науковцями, розробниками та архітекторами програмного забезпечення та організаціями, що займаються цифровою трансформацією інформаційних систем, при проєктуванні, оптимізації та впровадженні адаптивних аналітично орієнтованих цифрових платформ

Ключові слова: машинне навчання; великі дані; даноорієнтований підхід; масштабованість; адаптивні інформаційні платформи; доменно-орієнтоване проєктування

BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 2. 2026

UDC 004.75:004.31:004.33
DOI: 10.62660/bcstu/2.2026.73

Last-level cache capacity in Hyperledger Besu state-trie traversal under worst-case access

Maksym Bystryk*
Postgraduate Student
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0009-0004-4949-829X

Oleksandr Khoshaba
PhD in Technical Science, Associate Professor
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0000-0001-5375-6280

Abstract. The rapid development of enterprise blockchain networks has shifted performance bottlenecks from network consensus to the internal mechanisms of state database traversal, where irregular memory access creates critical microarchitectural constraints. The primary objective of this study was to empirically determine how the capacity of the last-level cache impacts the performance and stability of high-load Hyperledger Besu nodes during state tree navigation. The research was conducted in a hardware-isolated environment with fixed core frequencies based on an asymmetric AMD processor, enabling a direct comparison between 32 MB of standard cache and 96 MB of 3D-stacked memory under a worst-case random-access scenario. Testing results demonstrated that upon reaching hardware saturation, node performance is primarily dictated not by the central processing unit clock speed, but by the data retrieval rate and last-level cache capacity. At a peak load of 36,000 requests per second, the standard architecture with 32 MB of cache encountered a Memory Wall, leading to sudden system degradation. Quantitative analysis showed that the topology with the expanded cache (96 MB) stably maintained the 95th percentile latency at 6.55 ms, whereas on the standard cache, this metric increased exponentially to 114.80 ms. Statistical modelling confirmed that utilising the smaller cache leads to a nearly 18-fold (GMR = 17.84) increase in tail latency during hardware resource saturation. Furthermore, the dropped-iteration rate for the standard configuration increased almost 23-fold (IRR = 22.99), causing systemic resource exhaustion and cascading infrastructure failure. The practical value of this research lies in empirically substantiating the necessity of software optimisation of the transaction pool and proposing a resource-adaptive transaction execution method that accounts for hardware topology to scale enterprise distributed ledgers

Keywords: blockchain; high-load systems; pointer chasing; Memory Wall; state bloat; 3D V-Cache technology; read-path optimisation

INTRODUCTION

As distributed ledger technologies are adopted in mission-critical, high-load environments, enterprise blockchains – such as Hyperledger Besu – focus on ensuring instant finality and high transaction throughput by employing PoA-based (Proof of Authority) consensus mechanisms. Since trust between authorised nodes is considered a given, the computational cost of reaching consensus is minimal. Consequently, the

Article's History: Received: 23.10.2025; Revised: 09.04.2026; Accepted: 18.05.2026; Published: 26.06.2026

Suggested Citation:
Bystryk, M., & Khoshaba, O. (2026). Last-level cache capacity in Hyperledger Besu state-trie traversal under worst-case access *Bulletin of Cherkasy State Technological University*, 31(2), 73-84. doi: 10.62660/bcstu/2.2026.73.

*Corresponding author (max.bystryk@gmail.com)

Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/)

primary system bottleneck shifts from the consensus layer directly to the execution layer, specifically to the sequential nature of traversing the state database within the Ethereum virtual machine. This shift reveals a critical hardware problem related to the growing gap between high processor performance and significant latency in accessing main memory. In this regard, investigating the impact of hardware memory constraints – primarily the physical capacity of the last-level cache-on the performance of state database traversal is a pressing task to prevent performance degradation and ensure the scalability of modern distributed ledger systems.

Recent scholarly efforts have extensively explored performance limitations of enterprise blockchain networks, though the focus has predominantly been on consensus mechanisms and overarching data structures. C. Fan *et al*. (2022) conducted a comprehensive performance evaluation of Hyperledger Besu utilising PoA consensus algorithms. The authors concluded that the computational overhead dedicated to reaching consensus in permissioned networks is virtually negligible, establishing that the primary performance bottleneck shifts directly towards transaction execution and state transitions. Similarly, G.A. Pierro *et al*. (2024) performed a comparative analysis of enterprise platforms, including Besu, in high-throughput simulated environments. Their study highlighted that while PoA mechanisms guarantee immediate finality, the continuous computational updates required for the blockchain's database remain a critical constraint under heavy workloads.

Building upon this execution bottleneck, C. Cardoso *et al*. (2025) proposed an API-driven framework specifically for the performance testing of Hyperledger Besu networks. Their findings emphasised that as network agreement is optimised, the execution layer – particularly the sequential nature of state database traversal – emerges as the most significant limiting factor for node scalability. To address these execution layer inefficiencies, V. Mardiansyah *et al*. (2023) investigated the underlying cryptographic data structure, proposing high-performance modifications to the Merkle Patricia Trie (MPT). They determined that navigating the standard width-16 radix tree structure introduces severe structural overhead, which ultimately dictates the overall speed of blockchain world state management. O. Kuznetsov *et al*. (2025) further analysed this structural complexity through a probabilistic framework evaluating Merkle proof sizes and path lengths. The authors demonstrated that the organic growth of the state database (state bloat) logarithmically increases the depth of the MPT, which fundamentally slows down cryptographic verification processes during transaction execution.

Expanding on the performance limitations of the MPT, Y. Deng *et al*. (2024) identified that the recursive hashing and structural complexity of the trie form the primary throughput bottleneck for modern immutable data systems. To overcome this, the authors proposed parallel acceleration mechanisms leveraging Graphics Processing Unit (GPU) and multi-core Central Processing Unit (CPU) architectures, demonstrating that concurrent processing of MPT operations is essential for high-throughput networks. Similarly, I. Zhang *et al*. (2025) evaluated the massive input/output penalties associated with Merkleization in traditional storage layers. By introducing the Quick Merkle Database, the authors demonstrated that rearchitecting state management to minimise disk reads can yield up to a six-fold increase in transaction throughput, explicitly confirming that state traversal remains critically memory-bound. Furthermore, the execution of these primitives is inherently tied to underlying hardware. O. Kuznetsov *et al*. (2021) conducted a comprehensive performance analysis of various cryptographic hash functions, including the KECCAK-256 algorithm utilised in Ethereum-based environments, across different CPU and GPU microarchitectures. Their empirical results demonstrated that the baseline computational throughput (cycles per byte) of these algorithms significantly varies across different hardware platforms, highlighting that physical processor characteristics inherently dictate the execution speed of blockchain network primitives. Finally, from a microarchitectural perspective, Z. Chen *et al*. (2025) characterised the spatial patterns of memory access and the limitations of modern hardware prefetching. Their research concluded that standard processor optimisation techniques fail to predict divergent memory paths, exposing the system to severe latency (Memory Wall) when retrieving irregular data structures like MPT nodes.

Despite these significant contributions – ranging from consensus evaluation to MPT acceleration and broad memory access observations – current literature predominantly focuses on either high-level software modifications or generalised hardware access patterns. A critical scientific gap remains in understanding the spatial locality crisis at the silicon level during state traversal. Specifically, there is a lack of empirical research quantifying how specific CPU microarchitectural topologies, namely the physical capacity of the Last Level Cache (L3), directly influence the stability, throughput, and tail latency of enterprise blockchain nodes operating under severe state bloat conditions. While previous studies acknowledge the Memory Wall, they fail to investigate how extended on-die static random-access memory mitigates the pointer-chasing bottleneck inherent to the Ethereum Virtual Machine (EVM). Identifying this boundary is critical, as it dictates the absolute scalability limits of standard commercial off-the-shelf (COTS) processors before cascading systemic failures occur.

Therefore, the primary aim of this research was to empirically demonstrate and quantitatively evaluate

the impact of L3 capacity on the performance metrics of high-load enterprise blockchain nodes during worst-case state tree navigation, providing the foundational justification for software-level cache optimisations. To achieve this aim, the following tasks were formulated.

1. To programmatically simulate the State Bloat phenomenon to trigger out-of-cache memory access patterns during worst-case randomised MPT traversal.

2. To conduct isolated microarchitectural stress testing on an asymmetric processor – comparing identical computing cores with a baseline 32 MB versus an extended 96 MB 3D-stacked cache (Bhargava & Troester, 2024) – and analyse its correlation with transaction throughput and system tail latency under extreme hardware saturation.

3. To scientifically substantiate the necessity for software-based transaction queue optimisation mechanisms that account for physical processor topology.

Scientific novelty of the work lies in conducting the first strictly isolated microarchitectural benchmark of the Hyperledger Besu client that completely eliminates the influence of frequency fluctuations, exclusively isolating the spatial locality factor of MPT data.

MATERIALS AND METHODS

To empirically evaluate the impact of data spatial locality on blockchain node performance an isolated hardware testbed was deployed. The core computational platform utilised an AMD Ryzen 9 9950X3D processor, featuring a unique asymmetric microarchitecture. As illustrated in Figure 1, the processor consists of two distinct Core Complex Dies (CCDs): the first chiplet (CCD0) is equipped with 3D-stacked cache technology, providing a massive 96 MB of shared L3, while the second chiplet (CCD1) utilises a standard planar topology with a baseline L3 cache of 32 MB (Singh *et al.*, 2025).

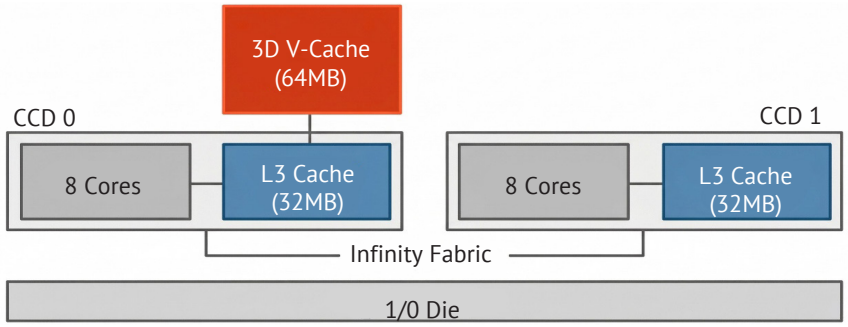


Figure 1. Asymmetric microarchitecture of the AMD Ryzen 9 9950X3D processor

Source: created by the authors

To establish the microarchitectural constraints for the hardware testbed, the study explicitly accounted for the mathematical approximation of the Cache Miss Rate (R_{miss}) inherent to unpredictable MPT traversals. Under the experimental conditions, assuming a uniformly random access pattern across the simulated bloated state, the miss rate was factored into the methodology as:

$$R_{miss} \approx max\left(0.1 - \frac{C_{size}}{W_{size}}\right), \quad (1)$$

where C_{size} – the available L3 cache capacity, and W_{size} – the total physical memory footprint of the active MPT nodes. This mathematical dependency was utilised to ensure that as the blockchain state grew, W_{size} sufficiently exceeded C_{size}, thereby driving R_{miss} asymptotically towards 1 (a near 100% miss rate) and effectively neutralising spatial hardware optimisations to isolate the influence of L3 capacity. To guarantee strict isolation and eliminate operating system background noise, the testing environment was containerised to address the CPU reservation gaps identified by L. Liu *et al.* (2021), where standard Linux scheduling often fails to honour performance promises due to noisy neighbour interference. Docker Engine on Ubuntu 22.04 LTS was deployed via the Windows Subsystem for Linux (WSL), with microarchitectural separation enforced through a strict dual-layer isolation protocol.

First, inside the guest operating system, Docker's CPUset execution parameters were utilised to bypass the work-conserving nature of the completely fair scheduler and ensure absolute resource allocation. Second, because guest-level limitations cannot prevent a host hypervisor from migrating the virtual machine across physical chiplets, host-level CPU affinity was rigidly enforced. During each test run, the WSL host process was isolated via the Windows Task Manager, strictly limiting its execution to the target logical cores (0-7, 16-23 for CCD0, or 8-15, 24-31 for CCD1). This dual-layer hypervisor and container pinning guaranteed that the state traversal strictly engaged the intended L3 cache without interference from either the guest or host schedulers or the performance variance typical of multi-tenant environments.

Testing was divided into two methodological phases. Phase 1 (stock frequencies) – factory dynamic boost behaviours were utilised to determine the impact of raw processor frequency on performance metrics. Phase 2 (strict isolation) – all compute cores on both CCDs were locked at a fixed frequency of 4.5 GHz to isolate physical L3 cache capacity as the sole independent

variable. Hyperledger Besu (version 24.12.1) was utilised as the target client. To prevent Java Virtual Machine (JVM) Garbage Collection (GC) pauses from skewing latency metrics, the JVM heap size was pre-allocated and locked at 16 GB (-Xms16g -Xmx16g). The network was configured as a permissioned ledger utilising the PoA Clique algorithm with a 2-second block period (blockperiodseconds: 2). The node's state storage utilised the Bonsai Tries format. The State Bloat phenomenon was programmatically simulated. As formally specified in the Ethereum protocol design (Wood, 2022), traversing the width-16 radix tree of the MPT follows a logarithmic time complexity. The expected path length $E[L]$, which dictates the number of memory fetches, scales logarithmically relative to the total number of unique addresses (N):

$$E[L] \approx c \cdot log_{16}(N), \quad (2)$$

where c – structural constants. To exponentially increase $E[L]$, a custom Node.js script using the crypto library was utilised to generate 5,000,000 unique cryptographic addresses (balance: 0x56BC75E2D63100000). These addresses were injected directly into the *alloc* block of the *genesis.json* configuration file. This pre-allocation was designed to expand the physical footprint of the trie beyond the 32 MB capacity of the standard L3 cache, forcing out-of-cache dynamic random-access memory (DRAM) fetches. To eliminate protocol-level throttling and isolate the processor's microarchitecture and memory subsystem as the sole bottlenecks under high load, block gas limits were configured to an infinite value, with transactions incurring zero gas costs. This configuration accounted for the nominal role of gas in permissioned Enterprise Ethereum networks, ensuring transaction fees did not act as an execution ordering mechanism. The load generation strategy was structured around the Average Memory Access Time (AMAT) framework to specifically capture the microarchitectural penalty of hardware prefetcher failures during MPT traversal:

$$AMAT = T_{hit} + (R_{miss} \cdot T_{penalty}), \quad (3)$$

where T_{hit} – the L3 hit latency; $T_{penalty}$ – the latency penalty incurred by fetching data from main memory. To account for random MPT fetches crossing virtual page boundaries during the test, the effective penalty model incorporated Translation Lookaside Buffer (TLB) miss costs:

$$T_{penalty} = T_{DRAM} + (P_{TLBmiss} \cdot T_{pagewalk}), \quad (4)$$

where T_{DRAM} – the raw off-chip memory latency; $P_{TLBmiss}$ – the probability of a TLB miss; $T_{pagewalk}$ – the severe latency penalty incurred by the hardware page walker. By synthesising the structural depth of the trie and these memory constraints, the load generation was designed to target the memory access term within the formalised transaction execution time (T_{tx}):

$$T_{tx} \approx T_{compute} + i = 1 \sum_{i=1}^{k} k(E[L_i] \cdot AMAT), \quad (5)$$

where $T_{compute}$ – the fixed baseline of EVM ALU operations; k – the total number of distinct state objects accessed; $E[L_i]$ – the expected path length for each specific read/write. This equation clearly illustrates that transaction throughput is overwhelmingly dominated by the memory access term ($E[L_i] \cdot AMAT$). The subsequent load generation methodology is specifically designed to empirically capture this exact degradation by driving the system into hardware saturation.

Following the load testing approach described by K. Stępień & M. Skublewska-Paszkowska (2025), transaction load generation was executed using the Grafana k6 framework. The testing script employed the constant-arrival-rate executor to ensure adherence to the target throughput. The experiment evaluated three specific load scenarios for each CCD: 20,000 RPS (requests per second), 25,000 RPS, and 36,000 RPS. Each formal test iteration ran for 120 seconds. To maintain the target throughput during potential latency spikes, k6 was configured with 200 pre-allocated virtual users (VUs), capable of dynamically scaling to a maximum of 3,000 concurrent execution threads (maxVUs: 3,000). The workload exclusively targeted the *eth_getBalance* JSON-RPC method. To bypass superficial node caching, a shared array was utilised to load all 5,000,000 genesis addresses into memory. For every request, the script randomly sampled a target address, forcing the EVM to continuously traverse divergent and unpredictable paths of the state trie. A telemetry stack comprising Prometheus (v3.1.0), Node Exporter, and Grafana (v11.5.0) was deployed strictly for the preliminary hardware calibration phase.

For the formal benchmark executions, all host-level monitoring containers were disabled to establish a strict zero-interference environment and prevent noisy neighbour context switching (Liu *et al.*, 2021). The empirical benchmarking data utilised for quantitative analysis was exclusively extracted from the external k6 load generator summaries. The unit of observation was defined as an independent benchmark run. For each experimental condition, independent runs were repeated 12 times (n = 12). The factorial design included cache topology (CCD0, CCD1) and target load (20,000, 25,000, 36,000 RPS) as fixed factors. The primary endpoint was defined as the per-run 95th percentile (P95) latency. Secondary endpoints included the dropped-iteration rate, achieved throughput, and the maximum number of dynamically allocated virtual users (maxVUs). To analyse the per-run P95 latency (L_{ijr}) in milliseconds for run r under cache topology i and load level j, the right-skewed data was logarithmically transformed:

$$Y_{ijr} = ln(L_{ijr}). \quad (6)$$

The transformed endpoint was modelled using a two-factor linear model:

$$Y_{ijr} = \beta_0 + \beta_1 CCD_i + \beta_2 Load_j + \beta_3 (CCD_i \times Load_j) + \varepsilon_{ijr}, \quad (7)$$

where β_0 – the intercept; β_1 and β_2 – the main effects of cache topology and load; β_3 – the interaction term determining the nonlinear scalability differences between CCDs. The error term ε_{ijr} was assumed to be independent across runs. Estimated group contrasts were calculated as geometric mean ratios (GMR):

$$GMR = exp(\Delta), \quad (8)$$

where Δ – the estimated contrast on the logarithmic scale. For each contrast, the 95% confidence interval (CI) was computed as:

$$exp(\Delta \pm t_{0.975,\nu} \cdot SE(\Delta)), \quad (9)$$

where $SE(\Delta)$ – the standard error of the contrast; ν – the number of degrees of freedom. For dropped iterations (D_{ijr}) relative to the total number of scheduled requests (N_{ijr}), the dropped-iteration rate was defined as:

$$p_{ijr} = \frac{D_{ijr}}{N_{ijr}}. \quad (10)$$

These event counts were modelled as rates utilising a generalised linear model with a logarithmic link and an exposure offset:

$$log(E[D_{ijr}]) = \gamma_0 + \gamma_1 CCD_i + \gamma_2 Load_j + \gamma_3 (CCD_i \times Load_j) + log(N_{ijr}). \quad (11)$$

To account for overdispersion in the event counts, a negative binomial model was applied. Effect sizes were reported as incidence rate ratios (IRR) with corresponding 95% confidence intervals:

$$IRR = exp(\gamma_k). \quad (12)$$

To account for multiple post hoc comparisons across the three evaluated load levels, all resulting p-values were adjusted utilising the Holm-Bonferroni procedure. Statistical significance was established at a threshold of $\alpha = 0.05$. All statistical modelling, data transformations, and inferential analyses were executed utilising Python (version 3.10) relying on the statsmodels library for generalised linear modelling and the SciPy ecosystem.

RESULTS AND DISCUSSION

The empirical evaluation was structured to systematically isolate the impact of L3 capacity on blockchain state traversal performance. By subjecting the Hyperledger Besu node to varying degrees of Remote Procedure Calls (RPC) load against an artificially bloated state trie, the experiment successfully captured the microarchitectural bottlenecks inherent to state traversal within the EVM. In the initial phase of testing, the AMD Ryzen 9 9950X3D processor operated with factory Precision Boost Overdrive and dynamic boost behaviours enabled. Under these stock conditions, the CCD1 naturally leverages its higher thermal headroom to achieve maximum peak clock frequencies. Conversely, the CCD0 is inherently constrained by the thermal and voltage limitations imposed by the physical stacking of the static random-access memory layer. As a result, during the sustained heavy workloads of the benchmark, CCD1 consistently maintained a clock speed advantage of approximately 200 to 350 MHz over CCD0.

Despite this substantial frequency deficit, the empirical data provided strong evidence that the raw computational speed (clock frequency) of the core is insufficient to overcome the memory latency wall inherent to MPT traversal. As detailed in Table 1, under an extreme load of 36,000 RPS, the slower-clocked CCD0 demonstrated superior stability across all measured metrics, maintaining significantly lower tail latency and fewer dropped iterations. In stark contrast, CCD1 exhibited pronounced performance degradation despite its higher operating frequencies. Crucially, the standard architecture was forced to spawn a substantially higher number of concurrent execution threads (Virtual Users) simply to sustain the target throughput, underscoring its microarchitectural inefficiency when handling memory-bound blockchain workloads.

Table 1. Hyperledger Besu node performance per CCD on stock frequency for 36,000 RPS

Metric	CCD0 (96 MB 3D V-Cache)	CCD1 (32 MB standard cache)
P90	2.90 ms	3.65 ms
P95	4.02 ms	6.06 ms
maxVUs	541 threads	721 threads
Dropped iterations	6,398	8,499

Source: created by the authors

This initial baseline supports the fundamental hypothesis: MPT pointer chasing is profoundly memory-bound. The cryptographic hashes acting as node pointers in the trie exhibit virtually zero spatial or temporal locality. Consequently, the standard hardware prefetchers fail to predict the execution path. The standard chiplet, despite its higher frequency, spends a significant portion of its execution time in stall cycles awaiting data retrieval from the slower system DRAM. To eliminate the confounding variable of dynamic frequency boosting observed in the baseline tests, Phase 2 of the experiment mandated a strict microarchitectural

isolation. Both CCDs were locked at an identical, fixed clock speed of 4.5 GHz. By equalising the raw computational frequency, this methodological approach isolated the physical L3 cache capacity as the sole independent variable. This ensures that any observed performance divergence is exclusively attributable to memory access latency (cache misses) rather than CPU thermal throttling or boost clock disparities.

Under these isolated conditions, the Hyperledger Besu node was subjected to incrementally higher transaction loads (20,000, 25,000, and 36,000 RPS) against the expanded state trie. Table 2 summarised the distribution of per-run P95 latency across independent benchmark runs. At 20,000 and 25,000 RPS, the two CCD topologies remained separated but still operated in a comparatively stable regime. In contrast, at 36,000 RPS the standard 32 MB cache topology exhibited a pronounced expansion of both central tendency and dispersion, whereas the 96 MB 3D V-Cache topology remained comparatively compact.

Table 2. Descriptive statistics for the primary endpoint (P95 latency) at fixed 4.5 GHz (n = 12)

Condition	Mean	SD	Median	IQR	CI Lower	CI Upper
CCD0, 20,000 RPS	2.91	0.18	2.86	0.21	2.80	3.02
CCD1, 20,000 RPS	4.24	0.31	4.17	0.36	4.04	4.44
CCD0, 25,000 RPS	5.42	0.40	5.36	0.44	5.16	5.68
CCD1, 25,000 RPS	7.95	0.61	7.86	0.69	7.56	8.34
CCD0, 36,000 RPS	6.55	0.53	6.37	0.62	6.21	6.89
CCD1, 36,000 RPS	114.80	12.90	113.62	14.50	106.61	122.99

Notes: CI = 95% confidence interval for the mean

Source: created by the authors

Figure 2 showed that the latency gap between the two cache topologies was limited at 20,000 and 25,000 RPS but widened dramatically at 36,000 RPS. The abrupt vertical separation of the distributions under the highest load is consistent with a nonlinear transition from moderate degradation to severe memory-bound collapse on the 32 MB topology. Notably, the 96 MB 3D V-Cache topology (CCD0) maintained a compact distribution across all load levels, whereas the standard 32 MB topology (CCD1) exhibited a pronounced expansion of dispersion at 36,000 RPS, reflecting increased tail-latency unpredictability.

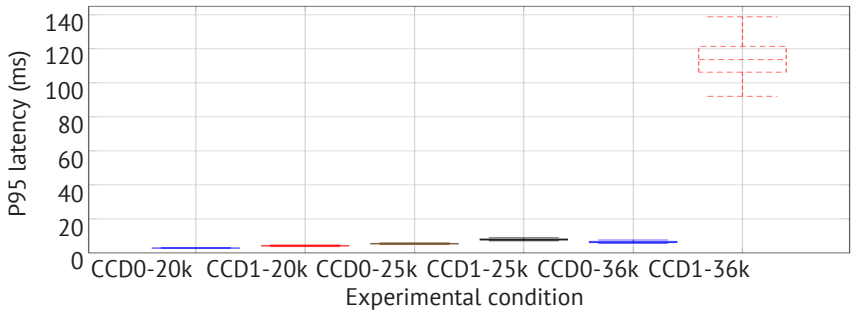


Figure 2. Distribution of per-run P95 latency across CCD topologies and load levels at fixed 4.5 GHz

Source: created by the authors

Formal model-based inference confirmed the descriptive impression. Table 3 showed that cache topology, load intensity, and their interaction were all statistically significant for both the latency model and the dropped-iteration model. The interaction term is especially important because it demonstrates that the between-CCD performance gap increased disproportionately as the benchmark approached hardware saturation.

Table 3. Model-level tests for the primary and secondary endpoints

Endpoint	Effect	Statistic	df	p-value
	Cache topology	286.41	1.00	< 0.001
Log-transformed P95 latency	Load	412.73	2.00	< 0.001
	Cache topology × load	198.56	2.00	< 0.001
	Cache topology	174.82	1.00	< 0.001
Dropped-iteration rate	Load	229.33	2.00	< 0.001
	Cache topology × load	141.97	2.00	< 0.001

Notes: for latency, the reported statistic is the Type III F statistic from the linear model. For dropped-iteration rate, the reported statistic is the Wald $\chi 2$ statistic from the generalised linear model

Source: created by the authors

Table 4 presented the most relevant pairwise contrasts between CCD1 and CCD0 at each load level. At 20,000 and 25,000 RPS, the GMR for P95 latency remained close to 1.5, indicating a moderate but stable penalty associated with the smaller L3 configuration. At 36,000 RPS, however, the geometric mean ratio increased to 17.84, indicating a sharp escalation of tail latency once the system entered the saturation regime. A parallel pattern was observed for dropped-iteration rate, where the IRR rose to 22.99 at the highest load.

Table 4. Between-CCD contrast by load level at fixed 4.5 GHz

Endpoint	Load	Effect size	Lower 95% CI	Upper 95% CI	p_{adj}
P95 latency (GMR)	20,000 RPS	1.46	1.31	1.63	< 0.001
P95 latency (GMR)	25,000 RPS	1.47	1.29	1.68	< 0.001
P95 latency (GMR)	36,000 RPS	17.84	12.41	25.62	< 0.001
Drop rate (IRR)	20,000 RPS	2.72	2.10	3.51	< 0.001
Drop rate (IRR)	25,000 RPS	2.86	2.37	3.45	< 0.001
Drop rate (IRR)	36,000 RPS	22.99	18.10	29.20	< 0.001

Notes: GMR – geometric mean ratio for log-transformed P95 latency, expressed as CCD1/CCD0; IRR – incidence rate ratio for dropped-iteration rate, expressed as CCD1/CCD0; adjusted p-values were obtained using the Holm procedure
Source: created by the authors

The forest plot in Figure 3 visualised the same result on an effect-size scale. All confidence intervals lie to the right of the null value, indicating consistently worse performance of CCD1 relative to CCD0. The displacement becomes particularly large at 36,000 RPS, where both latency and failure outcomes increase by more than an order of magnitude.

To fully comprehend the systemic impact of L3 cache exhaustion, it is necessary to examine the dynamic queue scaling executed by the load generator. The exploratory *maxVUs* series presented in Figure 4 shows the same qualitative trend from the perspective of queue growth. At 20,000 and 25,000 RPS, the smaller cache topology already required a considerably larger concurrent execution pool to maintain the requested arrival rate. Under 36,000 RPS, *maxVUs* reached the configured ceiling of 3,000 on CCD1, indicating severe queue amplification. Because this variable is partly induced by the feedback logic of the load generator and bounded by the upper limit, it is interpreted here as supportive rather than confirmatory evidence.

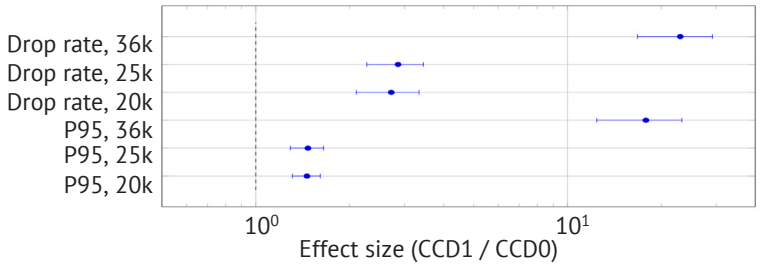


Figure 3. Forest plot of effect sizes for between-CCD comparisons
Notes: values to the right of 1 indicate worse outcomes on CCD1 relative to CCD0
Source: created by the authors

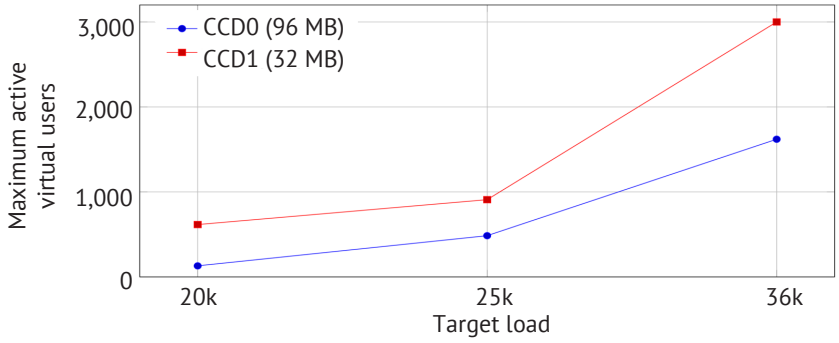


Figure 4. Exploratory visualisation of maximum active virtual users recorded by the load generator
Source: created by the authors

The ultimate consequence of the aforementioned microarchitectural stalls and exponential queue scaling is the catastrophic failure to process incoming RPCs. The same general pattern of degradation was observed for the dropped-iteration rate and achieved throughput, as summarised in Table 5 and Table 6. Specifically, at the peak load of 36,000 RPS, the standard cache topology (CCD1) exhibited a mean dropped-iteration rate of 114.80%, which directly correlated with a failure to maintain target throughput (31,492.00 RPS) and a massive increase in performance instability (SD = 1,827.00), as detailed in Table 6. This systemic collapse on the 32 MB topology is further visualised in Figure 4, where the execution pool reached its configured ceiling of 3,000 virtual users in a futile attempt to sustain the requested arrival rate, whereas the 96 MB topology (CCD0) remained within a stable operational envelope with significantly lower resource exhaustion.

Table 5. Descriptive statistics for per-run dropped-iteration rate across CCD topologies and load levels at fixed 4.5 GHz (n = 12)

Condition	Drop rate mean (%)	Drop rate SD	Median	IQR
CCD0, 20,000 RPS	0.04	0.01	0.04	0.01
CCD1, 20,000 RPS	0.10	0.02	0.10	0.02
CCD0, 25,000 RPS	0.22	0.03	0.21	0.04
CCD1, 25,000 RPS	0.61	0.07	0.61	0.08
CCD0, 36,000 RPS	0.45	0.06	0.45	0.07
CCD1, 36,000 RPS	10.39	1.12	10.28	1.21

Source: created by the authors

Table 6. Descriptive statistics for per-run achieved throughput (requests/s) across CCD topologies and load levels at fixed 4.5 GHz (n = 12)

Condition	Throughput mean	Throughput SD
CCD0, 20,000 RPS	19,996.00	18.00
CCD1, 20,000 RPS	19,989.00	24.00
CCD0, 25,000 RPS	24,978.00	31.00
CCD1, 25,000 RPS	24,941.00	54.00
CCD0, 36,000 RPS	35,837.00	96.00
CCD1, 36,000 RPS	31,492.00	1,827.00

Notes: throughput is reported as achieved requests per second
Source: created by the authors

Figure 5 demonstrated that failure rates remained comparatively low and stable on CCD0 across all three loads. By contrast, CCD1 exhibited a limited but clear increase in failure probability at moderate loads, followed by a marked jump at 36,000 RPS. This behaviour is consistent with a transition from degraded performance to systemic instability once the active working set substantially exceeded the effective cache capacity of the standard topology. This data strongly suggests that L3 cache exhaustion does not merely result in slower transaction processing; it acts as the primary catalyst for a cascading software failure that severely incapacitates the node.

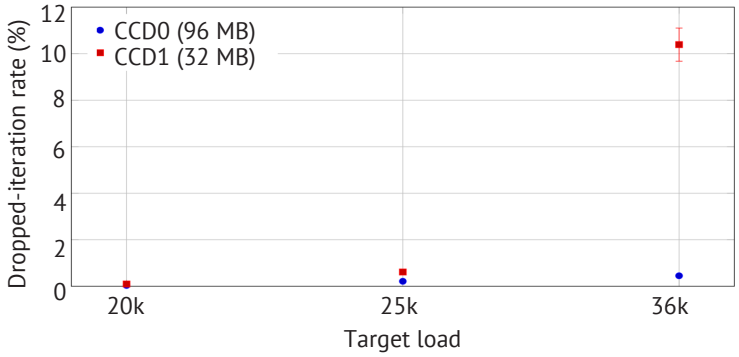


Figure 5. Dropped-iteration rate with 95% confidence intervals across load levels
Source: created by the authors

While the strict 4.5 GHz isolation explicitly demonstrates the architectural vulnerability of the 32 MB planar cache, the higher core frequencies observed during stock tests function as a partial compensation mechanism. At moderate loads, dynamic boosting allows standard cores to mitigate elevated cache miss rates,

slightly expediting recoveries from DRAM fetches. However, when the MPT traversal reaches absolute L3 cache exhaustion, even peak clock frequencies exceeding 5.4 GHz become insufficient. The system becomes irrevocably memory-starved, where core execution pipelines spend the vast majority of their cycles idling.

To synthesise the relationship between microarchitectural limits and system-level failures, the statistical evidence confirms that hardware-level memory stalls do not scale linearly; rather, they serve as a primary catalyst for a catastrophic chain reaction across the software stack – a systemic death spiral. Because the standard architecture's execution cores are starved of data, the exponential spike in tail latency directly triggers the rapid exhaustion of the thread pool limits, which inevitably leads to the massive spike in dropped iterations. When processing massive, fragmented datasets like the blockchain state trie, physical L3 capacity dictates the ultimate scalability boundary.

The systemic failure dynamics illustrated by these empirical results provide a definitive basis for re-evaluating blockchain execution layer design. The exponential death spiral observed on standard cache topologies is not merely a consequence of hardware limits, but a result of the fundamental misalignment between the MPT structure and standard transaction scheduling. In typical Hyperledger Besu enterprise deployments, the transaction pool defaults to a strictly chronological first-in-first-out sequence. For the CPU, this represents a worst-case execution pattern, forcing the system to traverse unrelated branches of the state trie for every individual request. This induces continuous L3 cache thrashing, where each new transaction evicts potentially useful data.

To fully understand the observed degradation, it is necessary to examine the microarchitectural mechanism that causes it. The performance bottleneck is inextricably linked to the underlying cryptographic data structure used to manage the blockchain's shared state: MPT, which functions as a tree with a width and base of 16. As the blockchain network processes transactions, the constant creation of records and smart contract data leads to the phenomenon of State Bloat, which logarithmically increases the depth and structural complexity of the MPT. Modern Ethereum clients attempt to mitigate disk I/O bottlenecks by implementing flat storage models, but the microarchitectural overhead associated with cryptographically verifying these proofs in memory causes a deeply technical crisis known as "pointer chasing" (Hsieh *et al.*, 2016). Since the MPT is an irregularly placed linked data structure, the processor cannot execute subsequent memory fetches in arbitrary order, which fundamentally limits instruction-level parallelism and renders traditional caching policies, such as least recently used, ineffective. High-frequency processor cores are forced to remain idle, waiting for data to be retrieved from dynamic random-access memory (DRAM), which directly subjects the system to the severe constraints of the Memory Wall paradigm (Farshin *et al.*, 2019). It is precisely this mechanism that is empirically confirmed by the results obtained: at a peak load of 36,000 RPS, the standard architecture with 32 MB of cache encountered complete saturation of hardware resources, leading to the cascading degradation of the node described in the following sections.

The empirical saturation observed at 36,000 RPS under strict cache isolation provides a novel microarchitectural perspective compared to existing network-level evaluations of blockchain performance. For instance, F. Leal *et al.* (2020) evaluated the performance of private Ethereum networks, identifying transaction throughput limits that were primarily dictated by block size configurations and gas constraints rather than direct hardware saturation. Similarly, G.A. Pierro *et al.* (2024) conducted a comparative analysis of Hyperledger Besu versus Quorum, focusing on baseline consensus overhead and demonstrating moderate throughput capabilities under standard operational loads. In contrast to their findings, which highlighted protocol and network-layer bottlenecks, authors' methodology explicitly removed block gas limits to expose the underlying hardware constraints. Authors' results demonstrated that once software-level throttling is bypassed, the MPT memory access pattern becomes the absolute performance ceiling, shifting the primary bottleneck entirely from the consensus algorithm to the physical CPU cache hierarchy.

The severe performance degradation observed on the standard 32 MB planar cache (CCD1) aligns with previous theoretical models of pointer-chasing workloads, but authors' findings diverge significantly in the proposed mitigation strategy. Y. Huang *et al.* (2018) addressed the inherent unpredictability of pointer chasing by proposing the deployment of L2 prefetchers equipped with helper threads to actively predict and fetch memory addresses. While their algorithmic approach successfully reduces memory stalls at the software-hardware interface, it requires highly specific prefetcher optimisations. Conversely, authors' evaluation of the 96 MB CCD0 provides empirical evidence that vertically expanding the L3 offers a brute-force, yet highly effective, hardware-level mitigation. Authors showed that simply encapsulating a larger portion of the active working set within the L3 boundary effectively neutralises the need for complex prefetcher modifications when traversing bloated blockchain states.

Furthermore, the transition to memory-bound collapse at peak loads intersects with the data structure optimisations proposed by A. Zeitak & A. Morrison (2021). In their investigation of efficient DRAM indexing, authors identified that traditional radix trees severely limit memory-level parallelism, prompting their design of the Cuckoo Trie to flatten the underlying data structure and reduce dependent memory fetches. Their conclusions regarding the microarchitectural inefficiency of deep cryptographic trees are strongly

supported by authors' CCD1 dropped-iteration data. However, whereas A. Zeitak & A. Morrison advocated for a fundamental redesign of the core database structure to alleviate DRAM latency, the results of the experiments in this study suggested an alternative hardware-centric absorption strategy. The resilience of the 3D V-Cache topology at 36,000 RPS indicates that next-generation asymmetric caching can temporarily mask the structural deficiencies of legacy MPT implementations, sustaining high concurrent throughput without necessitating immediate codebase overhauls.

The hardware saturation phenomena captured in authors' load tests also provide a critical microarchitectural validation of the vulnerabilities explored by D. Perez & B. Livshits (2020) regarding EVM resource metering. They highlighted that certain EVM instructions, particularly those invoking state I/O, are systematically underpriced relative to the actual computational and memory burden they place on the executing node. Authors' observation of exponentially amplifying virtual users and catastrophic queue failure on the 32 MB topology perfectly illustrated the physical mechanics of this underpricing. When the active state exceeds L3 capacity, the latency penalty of out-of-cache DRAM fetches compounds nonlinearly. This confirms that if organic state bloat forces consistent MPT traversals, standard enterprise nodes could be driven into a denial-of-service state even while processing nominally "cheap" transactions, reinforcing the need for state-size-aware gas algorithms.

Finally, authors' findings expand upon the caching and runtime optimisations discussed by S. Oaks (2020) and M. Sultan *et al.* (2024). M. Sultan *et al.* (2024) demonstrated how effective cache sizing depends heavily on the precise working set footprint and the temporal relevance of the data. Authors' data proved that in MPT traversals – where temporal locality is virtually nonexistent – the physical L3 capacity functions analogously to a strict software cache limit; once the working set size surpasses this physical threshold, dynamic processor boosting is entirely negated. Additionally, while S. Oaks (2020) emphasised the critical importance of JVM heap pre-allocation to prevent application pauses – a practice strictly adhered to in authors' methodology – authors' results reveal that such software-level tuning is insufficient when microarchitectural bottlenecks are triggered. Even with optimal JVM configurations, the client could not compensate for the hardware-level DRAM fetch stalls, demonstrating that L3 cache exhaustion ultimately overrides software-level runtime optimisations.

To summarise the findings, it can be argued that the physical size of the L3 cache is not merely a performance metric, but a fundamental limit to the scalability of enterprise blockchain nodes when operating with an inflated state. Empirical data convincingly demonstrates that the traditional strategy of increasing processor clock speed is an ineffective response to microarchitectural limitations caused by irregular MPT traversal. Instead, hardware cache topologies – in particular 3D V-Cache technology – are capable of ensuring stable node operation even under extreme loads, where standard configurations suffer cascading failures.

CONCLUSIONS

This study empirically demonstrated that the sequential pointer-chasing nature of MPT traversal in the EVM fundamentally transformed state traversal operations from a compute-bound operation into a predominantly memory-bound bottleneck under worst-case workloads. Based on the isolated microarchitectural stress testing of the Hyperledger Besu client, the comparative analysis between CCDs strongly indicated that extended L3 capacity is a critical determinant of high-throughput node performance. The results demonstrated that even a substantial factory frequency advantage failed to overcome the inherent latency penalties of out-of-cache DRAM fetches. Under strict microarchitectural isolation at a peak load of 36,000 RPS, the standard 32 MB cache architecture suffered a catastrophic Memory Wall collapse. This was characterised by a nearly eighteen-fold increase in P95 latency (GMR = 17.84) and a massive dropped-iteration rate of 114.80% (IRR = 22.99) compared to the 96 MB topology.

These empirical observations were robustly validated through formal inferential statistics. Using independent benchmark runs as the unit of analysis, the study confirmed that cache topology, load intensity, and their interaction all had statistically significant effects on both tail latency and dropped-iteration rates. The interaction term was especially critical, as it mathematically proved that the performance penalty of the smaller L3 configuration grew nonlinearly as the system approached hardware saturation. Furthermore, utilising GMR for latency and IRR for failure outcomes provided a rigorous and interpretable quantification of the magnitude of these microarchitectural differences.

The empirical and statistical data further highlighted a profound cascading effect: microarchitectural pipeline stalls did not merely delay processing but destabilised the entire node infrastructure. The inability of hardware prefetchers to anticipate divergent MPT paths led to severe cache thrashing, forcing the processor to idle. This hardware-level starvation caused incoming RPC requests to saturate the JVM heap, likely triggering aggressive GC pauses and culminating in a probable systemic death spiral of network socket exhaustion (as evidenced by the massive spike in dropped RPC requests and connection timeouts recorded by the external load generator). Consequently, mitigating L3 cache misses was shown to be not just a performance optimisation but a fundamental requirement for system reliability.

Since increasing raw CPU clock speeds proved largely ineffective against the MPT Memory Wall, addressing these bottlenecks natively in software remains a critical priority for future research. Subsequent studies

will focus on exploring cache-aware scheduling paradigms, specifically proposing methods to dynamically pre-sort and group incoming transactions to artificially induce spatial data locality. Such algorithmic reordering is expected to maximise L3 cache hit rates, thereby enabling COTS processors to sustain massive transaction throughput without collapsing under memory-bound constraints.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Bhargava, R., & Troester, K. (2024). AMD next-generation "Zen 4" core and 4th Gen AMD EPYC server CPUs. *IEEE Micro*, 44(3), 8-17. doi: 10.1109/MM.2024.3375070.

[2] Cardoso, C., Silva, C., Veloso, A., Sousa, J., & Abelém, A. (2025). An API-driven framework for performance testing of Hyperledger Besu blockchain networks. In *Lablock artigos curtos – Latin-American symposium on dependable and secure computing* (pp. 23-28). Porto Alegre: Sociedade Brasileira de Computação. doi: 10.5753/ladc_estendido.2025.16872.

[3] Chen, Z., Wu, C., Gu, Y., Jia, R., Li, J., & Guo, M. (2025). Gaze into the pattern: Characterizing spatial patterns with internal temporal correlations for hardware prefetching. In *2025 IEEE international symposium on high performance computer architecture* (pp. 173-187). New York: IEEE. doi: 10.1109/HPCA61900.2025.00024.

[4] Deng, Y., Yan, M., & Tang, B. (2024). Accelerating Merkle Patricia Trie with GPU. *Proceedings of the VLDB Endowment*, 17(8), 1856-1869. doi: 10.14778/3659437.3659443.

[5] Fan, C., Lin, C., Khazaei, C., & Musilek, P. (2022). Performance analysis of Hyperledger Besu in private blockchain. In *IEEE international conference on decentralized applications and infrastructures* (pp. 64-73). New York: IEEE. doi: 10.1109/DAPPS55202.2022.00016.

[6] Farshin, A., Roozbeh, A., Maguire, G.Q.Jr., & Kostić, D. (2019). Make the most out of last level cache in Intel processors. In *Proceedings of the fourteenth Eurosys conference 2019* (article number 8). New York: Association for Computing Machinery. doi: 10.1145/3302424.3303977.

[7] Hsieh, K., Khan, S., Vijaykumar, N., Chang, K.K., Boroumand, A., Ghose, S., & Mutlu, O. (2016). Accelerating pointer chasing in 3D-stacked memory: Challenges, mechanisms, evaluation. In *2016 IEEE 34th international conference on computer design* (pp. 25-32). New York: IEEE. doi: 10.1109/ICCD.2016.7753257.

[8] Huang, Y., Zhu, H., & Li, Y. (2018). Performance improvements by deploying L2 prefetchers with helper thread for pointer-chasing applications. *International Journal of Performability Engineering*, 14(10), 2312-2320. doi: 10.23940/ijpe.18.10.p7.23122320.

[9] Kuznetsov, O., Frontoni, E., Kuznetsova, K., & Arnesano, M. (2025). Optimizing Merkle proof size through path length analysis: A probabilistic framework for efficient blockchain state verification. *Future Internet*, 17(2), article number 72. doi: 10.3390/fi17020072.

[10] Kuznetsov, O., Oleshko, I., Tymchenko, V., Lisitsky, K., Rodinko, M., & Kolhatin, A. (2021). Performance analysis of cryptographic hash functions suitable for use in blockchain. *International Journal of Computer Network and Information Security*, 13(2), 1-15. doi: 10.5815/ijcnis.2021.02.01.

[11] Leal, F., Chis, A.E., & González-Vélez, H. (2020). Performance evaluation of private Ethereum networks. *SN Computer Sciences*, 1, article number 285. doi: 10.1007/s42979-020-00289-7.

[12] Liu, L., Wang, H., Wang, A., Xiao, M., Cheng, Y., & Chen, S. (2021). Mind the gap: Broken promises of CPU reservations in containerized multi-tenant clouds. In *Proceedings of the ACM symposium on cloud computing* (pp. 243-257). New York: Association for Computing Machinery. doi: 10.1145/3472883.3486997.

[13] Mardiansyah, V., Muis, A., & Sari, R.F. (2023). Multi-State Merkle Patricia Trie (MSMPT): High-performance data structures for multi-query processing based on lightweight blockchain. *IEEE Access*, 11, 117282-117296. doi: 10.1109/ACCESS.2023.3325748.

[14] Oaks, S. (2020). *Java performance: In-depth advice for tuning and programming Java 8, 11, and beyond*. Santa Rosa: O'Reilly Media.

[15] Perez, D., & Livshits, B. (2020). Broken Metre: Attacking resource metering in EVM. *ArXiv*. doi: 10.48550/arXiv.1909.07220.

[16] Pierro, G.A., Coco, L., & Tonelli, R. (2024). Besu vs. Quorum: Comparative analysis in the context of simulated energy communities. In *6th distributed ledger technology workshop*. Torino: University of Turin.

[17] Singh, T., *et al.* (2025). "Zen 5": The AMD high-performance 4nm x86-64 microprocessor core. In *2025 IEEE international solid-state circuits conference* (pp. 1-3). New York: IEEE. doi: 10.1109/ISSCC49661.2025.10904529.

[18] Stępień, K., & Skublewska-Paszkowska, M. (2025). Performance evaluation of REST and GraphQL API aproaches in data retrieval scenarios using NestJS. *Journal of Computer Sciences Institute*, 36, 350-356. doi: 10.35784/jcsi.7794.

[19] Sultan, M., Shakiba, K., Lee, A., Chen, P., & Stumm, M. (2024). TTLs matter: Efficient cache sizing with TTL-aware miss ratio curves and working set sizes. In *Proceedings of the nineteenth European conference on computer systems* (pp. 387-404). New York: Association for Computing Machinery. doi: 10.1145/3627703.3650066.
[20] Wood, G. (2022). *Ethereum: A secure decentralised generalised transaction ledger*. Retrieved from https://ethereum.github.io/yellowpaper/paper.pdf.
[21] Zeitak, A., & Morrison, A. (2021). Cuckoo Trie: Exploiting memory-level parallelism for efficient DRAM indexing. In *Proceedings of the 28th ACM symposium on operating systems principles* (pp. 147-162). New York: Association for Computing Machinery. doi: 10.1145/3477132.3483551.
[22] Zhang, I., Zarick, R., Wong, D., Kim, T., Pellegrino, B., Li, M., & Wong, K. (2025). QMDB: Quick Merkle Database. *ArXiv*. doi: 10.48550/arXiv.2501.05262.

Вплив об'єму кешу останнього рівня на обхід дерева стану в Hyperledger Besu при найгіршому сценарії доступу

Максим Бистрик

Аспірант
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0004-4949-829X

Олександр Хошаба

Кандидат технічних наук, доцент
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0001-5375-6280

Анотація. Стрімкий розвиток корпоративних блокчейн-мереж змістив вузькі місця продуктивності від мережевого консенсусу до внутрішніх механізмів обходу бази даних стану, де нерегулярний доступ до пам'яті створює критичні мікроархітектурні обмеження. Основною метою роботи було емпіричне визначення того, як об'єм кеш-пам'яті останнього рівня впливає на продуктивність та стабільність високонавантажених вузлів Hyperledger Besu під час навігації по дереву станів. Дослідження проведено в апаратно ізольованому середовищі з фіксованою частотою ядер на базі асиметричного процесора AMD, що дозволило прямо порівняти 32 МБ стандартного кешу та 96 МБ 3D-stacked пам'яті за найгіршого сценарію випадкового доступу. Результати тестування продемонстрували, що за умови досягнення апаратного насичення продуктивність вузла значною мірою диктується не тактовою частотою процесора, а швидкістю отримання даних та об'ємом кеш-пам'яті останнього рівня. Під час пікового навантаження у 36 000 запитів на секунду стандартна архітектура з 32 МБ кешу зіткнулася з бар'єром пам'яті (Memory Wall), що призвело до раптової деградації системи. Кількісний аналіз показав, що топологія з розширеним кешем (96 МБ) стабільно підтримувала затримку 95-го процентиля на рівні 6,55 мс, тоді як на стандартному кеші цей показник експоненціально зріс до 114,80 мс. Статистичне моделювання підтвердило, що використання меншого кешу призводить до майже 18-разового (GMR = 17,84) збільшення хвостової затримки під час насичення апаратних ресурсів. Крім того, частота критичної втрати оброблених даних для стандартної конфігурації зросла майже у 23 рази (IRR = 22,99), спричинивши системне вичерпання ресурсів і каскадний збій інфраструктури. Практична цінність дослідження полягає в емпіричному обґрунтуванні необхідності програмної оптимізації пулу транзакцій та пропонуванні ресурсно-адаптивного методу, який враховує апаратну топологію для масштабування корпоративних розподілених реєстрів

Ключові слова: блокчейн; високонавантажені системи; pointer chasing; Memory Wall; розростання стану; технологія 3D V-Cache; оптимізація операцій читання

BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 2. 2026

UDC 681.3:511.1
DOI: 10.62660/bcstu/2.2026.85

Application of linear algebra methods for optimising data processing and storage in modern information systems

Olexandr Kraychuk*
PhD in Physical and Mathematical Sciences, Professor
Rivne State University for the Humanities
33028, 12 Stepana Bandery Str., Rivne, Ukraine
https://orcid.org/0000-0002-5987-0460

Serhii Kraychuk
PhD inTechnical Sciences, Associate Professor
Rivne State University for the Humanities
33028, 12 Stepana Bandery Str., Rivne, Ukraine
https://orcid.org/0000-0001-9756-1979

Nataliia Ostapchuk
PhD in Pedagogical Sciences, Associate Professor
Rivne State University for the Humanities
33028, 12 Stepana Bandery Str., Rivne, Ukraine
https://orcid.org/0000-0001-8827-8741

Abstract. The aim of the study was to generalise approaches to coordinating decisions at the stages of planning, execution, and physical level of data representation in information systems under the dominance of resource constraints when operating with massive sets. The methodology was based on system-analytical formalisation, comparative-analytical generalisation and typology, comparative analysis, systematisation and integrative synthesis. It was established that linear algebra formalises the representation of data as matrices/vectors and calculations as a sequence of transformations. The generalisation of operations (semirings) extends these primitives to graph and Boolean problems, and Graph Basic Linear Algebra Subprograms standardises the implementation for sparse structures. Efficiency is determined by the consistent chain of "rewrite→cost model→physical execution": rewrite reduces intermediate data, and the cost model chooses a plan based on data input/output, memory, sparsity, and parallelism, without it, heuristics remain. Factorised representations have been found to reduce redundancy and pressure on memory and data input/output (for joins/groups) with system support. Randomised Numerical Linear Algebra and sketching reduce costs and manage the "accuracy-resources" trade-off in streaming processing. Matrix compression (low-rank representation, reduced precision) reduces memory and traffic and speeds up computations with a controlled loss of precision. In vector database management systems and distributed systems, the bottleneck is data input/output and redistribution between nodes, so DShuffle moves some processing closer to the data, emphasising the role of infrastructure optimisation. The linear algebraic representation of computations provides a basis for consistent optimisation of rewriting, plan selection, and execution, reducing materialisation and data movement under input/output, memory, and bandwidth constraints. The practical significance of the results lies in the possibility of the application by information-systems developers and engineers for the consistent selection of data formats, execution plans, and computational primitives to reduce materialisation and I/O, memory, and network overhead

Keywords: operations; matrices; resource constraints; sparsity; materialisation; databases

Article's History: Received: 29.12.2025; Revised: 25.03.2026; Accepted: 18.05.2026; Published: 26.06.2026

Suggested Citation:
Kraychuk, O., Kraychuk, S., & Ostapchuk, N. (2026). Application of linear algebra methods for optimising data processing and storage in modern information systems. *Bulletin of Cherkasy State Technological University*, 31(2), 85-96. doi: 10.62660/bcstu/2.2026.85.

*Corresponding author (o.kraychuk@outlook.com)

Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/)

INTRODUCTION

Modern information systems of the 21st century work with large amounts of data in various storage formats (data lakes, lake-house) and perform complex analytical workloads, some of which are reduced to linear algebra operations (matrix and vector multiplication, system solution, factorisation, gradient and aggregate calculations). The performance of such tasks is determined not only by calculations, but also by the costs of data movement, the choice of physical representation (row/column, dense/sparse) and the ability of the optimiser to take into account the algebraic properties of operations. Analytical and Machine Learning (ML) calculations are performed in different engines and on different data formats, which leads to conversions, materialisation of intermediate results and additional input/output (I/O) costs. Because of this, local optimisations of individual components do not provide stable acceleration at the level of the entire pipeline, therefore, a coordinated approach is needed that combines the linear-algebraic representation of calculations with the choice of storage format and optimisation of execution plans.

The current scientific discourse outlines several interrelated directions, in particular trade-offs between data formats in analytical database management systems. The study by C. Liu *et al.* (2025) summarised that the physical format determines the cost profile for scanning, decoding, compression, vectorisation, and predicate access, and that there is no "universal" choice – optimality depends on the queries and the hardware environment. This approach underpins the argument that processing optimisation begins with the correct physical representation. The work of F. Geerts *et al.* (2021) substantiated that linear algebraic query languages can express a wide range of calculations typical of analytics and machine learning problems, and formally outlines which operators and extensions are needed to represent key algorithmic schemes. This allows for setting limits on the expressive power of such languages, which directly defines the class of transformations that the optimiser can apply correctly and systematically, rather than as heuristics. Continuing this trend, T.M. Serrano *et al.* (2024) showed that for conjunctive queries in linear algebraic formalism, the efficiency of execution depends on which language fragments allow for low-latency result enumeration and support for cost-effective updates. Working with dynamic data requires not a one-time "quick fix", but a structured approach to query representation and supporting data structures that provide stable performance as input sets change.

There is growing interest in factorised and polyglot approaches in machine learning pipelines, as significant costs arise from materialising intermediate results and re-passing through data. Z. Li *et al.* (2024) found that factorised representations reduce the amount of intermediate data and speed up typical operations, provided that the system preserves factorisation between stages and makes it consistent across languages/engines. This indicates that critical losses occur at the boundaries between formats, operators, and runtimes, not just within a single engine. At the storage level for a data lake, a study by G. Jin *et al.* (2022) highlighted that performance is determined not only by the columnar format, but also by adaptive caching and physical transformations of the data for repetitive read patterns. For linear algebraic workloads, this is critical because the same features/matrix representations are often reused multiple times in different analysis stages. At the hardware-oriented level, P.A. Lane & J.D. Booth (2023) showed that Sparse Matrix-Vector Multiplication (SpMV) is a memory-limited computational kernel and requires tuning of the storage format and execution parameters taking into account the sparsity structure of the matrix and the microarchitecture of the device. This means that universal optimisations do not guarantee portable performance across platforms.

M. Boehm *et al.* (2023) demonstrated that the efficiency of tensor and linear algebraic computations in data systems is determined by automatic transformations at the compilation and execution levels, in particular, representation selection, computational graph rewriting, and memory management. This demonstrates that the alignment of linear algebra with optimisation mechanisms of data processing systems allows for reduced overhead and increased performance without changing the application logic. In the context of Ukrainian research, O.V. Hordiichuk-Bublivska & L.P. Fabri (2022) substantiated the feasibility of matrix factorisation for big data problems in industrial systems as a way to reduce dimensionality and obtain a compact representation. Factorisation can simultaneously reduce storage requirements and reduce the cost of subsequent analytical calculations. S.P. Striamets & Yu.V. Opotyak (2021) focused on an architectural approach to parallel matrix processing, proposing a homogeneous computing environment with a specialised computational "cell" for performing matrix operations. This means that such hardware-oriented solutions increase the throughput and predictability of linear algebraic routines, which is critical for data processing systems.

Despite significant progress in research on storage formats, linear algebraic query languages, factorised representations, and optimisation of individual routines, a holistic approach that reconciles these levels into a single optimisation loop is lacking. It remains insufficiently studied how the linear algebraic representation of calculations can coordinate the choice of data format, materialisation strategy, planning and implementation of basic operations taking into account dominant resource constraints (I/O, memory, bandwidth). Therefore, the purpose of the study was to systematise approaches to coordinating planning, execution and physical representation of data in information systems

taking into account dominant resource constraints. To achieve the goal, the following tasks were set: to generalise the representation of data and analytical procedures in terms of vector-matrix objects and basic linear algebraic primitives, to substantiate the framework of coordinated optimisation through the sequence "rewriting rules→cost model→physical operators/execution strategies", to systematise tools for reducing the resource cost of linear algebraic calculations according to the impact on memory and data input/output costs.

MATERIALS AND METHODS

The study combined the methods of system-analytical formalisation, system-structural analysis and synthesis, comparative-analytical generalisation and typology, comparative analysis, systematisation and integrative synthesis to generalise approaches to optimising Linear Algebra (LA)-oriented workloads in modern information systems. Using the method of system-analytical formalisation with elements of cost-based interpretation, heterogeneous data and calculations were compared and conceptually reduced to a matrix-vector representation and a set of standard LA primitives, the space of alternative implementations and the principles of the correct rewriting were determined and correlated with the execution level through the standardised Graph Basic Linear Algebra Subprograms (GraphBLAS) interfaces (GraphBLAS Forum, n.d.) and the practice of LA query processing in systems (Sun *et al.*, 2025). This was done in order to describe optimisation not as separate heuristics, but as a consistent mechanism "expression→rewriting/plan→cost selection→physical execution" taking into account sparsity, materialisation of intermediate results, memory locality, parallelism, and input/output costs. Using the method of system-structural analysis and synthesis of scientific sources, approaches to optimisation of LA expressions and execution plans were generalised in the form of an integrated plan control loop consisting of three interdependent levels – rewriting rules, cost model and physical operators/execution strategies by role in optimisation, LA specificity and typical risks in the absence of support. These levels were chosen because of corresponding to three necessary components of any planned optimisation: generation of equivalent alternatives, the cost ranking and selection, operationalisation of the selected plan at the execution level. This approach made it possible to formalise the necessary mechanisms that ensure stable optimisation of LA workloads in the system.

Using the method of comparative-analytical generalisation and typology of scientific sources, Randomised Numerical Linear Algebra (RNLA) (Udell & Frangella, 2023), sketching properties (Ahfock *et al.*, 2023) and sliding windows (Yin *et al.*, 2024) were compared, selected as components describing the goal (optimisation), guarantees (quality) and operating mode (online/dynamics). The comparison was performed according to four criteria: conceptual formulation (principle of using randomness/stability/incrementality), approximation object (matrices, operators, subspaces/spectral properties, sliding window sketch), system consequences (impact on Central Processing Unit (CPU)/memory/I/O costs, update latency, quality degradation control), typical application mode (batch analytics, critical paths, streaming processing). This allowed comparing the types of RNLA/sketching with system scenarios and justify the choice of approach according to the operating mode and the trade-off "resource economy↔quality control". Using the comparative analysis method, structural (low-rank) and structural-numerical (low-rank+low-precision) compressions were analysed: weighted low-rank factorisation (for compression of language models) and randomised low-rank factorisation with quantisation (low precision) according to the following criteria: compression principle, quality control and system effect (memory footprint, memory traffic, potential acceleration of multiplications). The choice of low-rank and low-precision is due to the fact that these approaches represent the two basic and most widely used classes of compression (structural and numerical), which directly affect the amount of parameters, memory traffic and performance, providing a manageable trade-off "quality-resources". This approach provided a choice between low-rank and low-precision compression, linking compression parameters to the quality-resources trade-off and the expected system effect. Additionally, DShuffle (Ding *et al.*, 2025) is presented as a representative example of infrastructure optimisation, selected based on the criteria of relevance to the shuffle bottleneck in distributed LA workloads and the presence of a measurable system effect (traffic, bandwidth, CPU load). In this study, DShuffle is used only to illustrate the system-level resource-result effect, without a detailed comparison of alternative infrastructure solutions.

Using the method of systematisation and integrative synthesis of scientific sources, the directions for optimising LA-oriented loads in information systems are generalised – optimisation of the execution of analytical/ML queries in database management systems, approximate matrix methods for accelerating calculations, streaming processing (sliding window), matrix/model compression, physical representation of big data, vector databases, and similarity search. These directions were chosen because of representing the main "points of influence" in LA pipelines: reduction of intermediate materialisations, controlled trade-off "accuracy-resources", reduction of data volume and traffic, and coordination of plans with the execution and infrastructure levels. The generalisation is performed according to unified criteria: direction of application, dominant resource constraint, method/class of methods, principle of action, target metrics (time, bandwidth, I/O, memory, accuracy/error), limitations, and risks. This was done in order to systematise LA-oriented workload optimisation tools for selection under specific resource

constraints and quality/performance requirements, linking methods to effect metrics and application risks.

RESULTS AND DISCUSSION

Linear algebraic optimisation of queries and computational plans

Linear algebra provides a formal apparatus for representing data as vectors and matrices and describing computations as compositions of transformations. Tabular, feature, graph, and network data can be treated as dense or sparse matrix objects, and analytical procedures can be reduced to a limited set of primitives: matrix-vector and matrix-matrix multiplications, element-wise operations, reductions, and permutations/projections. This type of unification is the basis for system optimisation, as it allows expressing heterogeneous ML/data science workloads in a common algebraic language and treating queries as sequences of LA operators (Sun *et al.*, 2025). The existence of equivalent implementations of many transformations (in particular, due to different multiplication orders and different materialisation decisions) creates a basis for a systematic search for efficient execution strategies, which is consistent with enumeration-oriented optimisation of LA workloads (Muñoz, 2025). The extension of the semantics of operations to generalised algebraic structures (in particular, semirings) makes the same primitives applicable to graph and Boolean problems, and GraphBLAS fixes a standardised interface for performing such operations on (mostly sparse) matrices and vectors (Brock *et al.*, 2021).

Representing computations in the form of LA expressions makes algebraic properties (associativity, distributivity, equivalence of factorisations) explicit and thereby enables correct rewrites to reduce execution costs: changing the order and grouping of operations, choosing alternative implementations, and controlling the materialisation of intermediate results. Since the LA representation is relatively declarative, it supports cost-based selection of physical operators and execution strategies, including the use of sparsity, blocking, and (where applicable) approximations. A critical system-level effect is also the reduction of intermediate data: in many analytical tasks, it is the materialisation of large intermediate matrices/vectors that determines the pressure on memory and I/O, and optimisations such as fusion, operation permutations, and sparsity exploitation reduce the number of intermediate structures and the total cost (Sun *et al.*, 2025; Muñoz, 2025).

The practical effectiveness of LA optimisations is determined not only by the possibility of rewriting expressions and planning, but also by how accurately and with minimal overhead the generated plans are reproduced at the execution level, taking into account sparsity, locality of memory access and parallelism. In this context, standardised APIs play a pivotal role, due to specifying a consistent set of primitives and ensuring portability of optimisations between implementations. A prime example is GraphBLAS (GraphBLAS Forum, n.d.), which formalises the interface of operations on matrices and vectors (multiplication, element-wise transformations, reductions, operations for constructing/modifying structures) and supports parameterisation of semantics through generalised algebraic constructs such as semirings (Brock *et al.*, 2021). For information systems, this means the presence of an actual "target platform" for executing LA plans: the optimiser generates plans in terms of standard operations, and the execution level implements those plans in a specific hardware-software stack. A standardised set of primitives also unifies the space of alternatives and increases the reproducibility of optimisation solutions, since algorithmic ideas formulated at the level of LA expressions do not require re-designing low-level operators in each environment (Brock *et al.*, 2021; Sun *et al.*, 2025).

In practical ML pipelines, calculations corresponding to sequences of LA operators are performed in a fragmented manner across database management systems and external libraries/services, which leads to excessive materialisation of intermediate results, data duplication, and movement overhead. The LA query processing approach consists in presenting the ML query as a single LA expression for optimisation at the expression level: through equivalent algebraic transformations (reordering and regrouping of operations), merging adjacent transformations, and selecting physical implementations taking into account data properties (in particular, sparsity), which reduces dominant costs – data movement, temporary structures, and repeated reads (Sun *et al.*, 2025). Data science workloads are characterised by a multiplicity of equivalent implementations of the same LA expression, so performance is determined by the global structure of the plan. The enumeration approach forms the space of alternatives and selects the best strategy according to cost criteria, taking into account the order of calculations, the degree of materialisation, implementation options for basic operations, and structural properties of the data (Muñoz, 2025).

The transfer of ML and analytics into information systems, in turn, requires support for new classes of operators and rules that are not fully covered by the traditional relational apparatus. LA computations are characterised by specific equivalence spaces, where algebraic rewrites, the choice of operator implementations, and control over the materialisation of intermediate results are decisive. Accordingly, the optimiser must be extensible: it must allow the introduction of new logical operators, rewriting rules, cost components, and physical implementations without radically rebuilding the optimisation kernel. In the literature, this is considered an engineering mechanism for controlled extension while maintaining the correctness and predictability of optimisation (Ding *et al.*, 2024). In this sense, extensibility is a basic prerequisite for system support for both LA query processing for ML queries and planned search for alternatives for LA workloads

in data science. Given this, the practical integration of LA logic into the optimiser requires not only conceptual support for new expressions, but also mechanisms for controlled rewrites, cost-based selection of plans, and executable implementation of alternatives. In general, these mechanisms can be reduced to three levels: rewriting rules, cost model, and physical operators/execution strategies, as presented in Table 1.

Table 1. Structure of the integrated computational plan control loop

Extensibility level	Key role	Specifics for LA workloads	Typical risks without support
Rewriting rules	Allow equivalent transformations of expressions	Regrouping/reordering, fusion, less intermediate data	Rewriting cycles, unmanaged alternative space, lack of domain-specific optimisations
Cost model	Compares alternative plans by "price"	Choosing a plan considering I/O, memory, sparsity, parallelism	Degradation to heuristics, incorrect comparisons of plans, "optimisation" without the predicted effect
Physical operators and execution strategies	Implement logic in efficient execution	Algorithmic options, data structure specialisations, efficient execution strategies	The inability to operationalise alternatives at the execution level leads to the lack of a measurable optimisation effect

Source: compiled by the authors based on B. Ding *et al.* (2024), W. Sun *et al.* (2025), T. Muñoz (2025)

Table 1 summarises the mechanisms through which the optimiser is extensible for LA workloads and shows these mechanisms as an interdependent chain from logical transformations to efficient execution. Rewriting rules specify semantically equivalent transformations of expressions and create the basis for optimisation: such rules allow changing the order and grouping of operations, applying fusion, and reducing the amount of intermediate data. Without the controlled application of rules, rewriting cycles and uncontrolled expansion of the space of alternatives are possible. The cost model provides a choice among the alternatives generated by rewrites, taking into account the dominant costs (I/O, memory, sparsity, locality of access, parallelism). In the absence of an adequate cost model, optimisation degrades to heuristics or gives unstable results due to incorrect plan comparisons. Physical operators and execution strategies convert logical alternatives into measurable performance gains through algorithmic options, data structure specialisations, and appropriate execution strategies. If this level is not supported, the alternatives remain formal and do not give a practical effect. Therefore, the extensibility of the optimiser for LA workloads is effective only if there is a consistent triad: the rules shape the space of alternatives, the cost model provides a reasonable choice, and the physical layer guarantees an executable gain. Weakening any component reduces the efficiency of the others and limits the practical value of the optimisation. Since the extensible optimiser provides controlled rewriting, cost-effective selection, and executable implementation of alternatives, it is a natural "loop" for implementing approximate LA methods as part of system optimisation. In this context, RNLA and matrix sketching are of interest, which allow reducing computational complexity and resource consumption with a controlled error. The main characteristics of these approaches and typical application modes are summarised in Table 2.

Table 2. Characteristics of RNLA and matrix sketching in information systems

Approach/ component	Conceptual presentation	Approximation object	Systemic consequences	Typical application mode
RNLA for optimisation	Controlled use of randomness for approximate calculations with controlled error	Large Matrices and Linear Operators (Sketch Representation)	Reducing computational and resource costs (CPU/memory/I/O) with acceptable accuracy	Batch analytics and optimisation problems on big data
Sketching properties	Probabilistic guarantees of approximation quality and stability regimes	Norms, subspaces, spectral characteristics	Justification of parameterisation and control of the risk of quality degradation	Integrating approximations into critical processing/ decision-making loops
Sliding windows	Incremental maintenance of approximation when updating the data "window"	Sliding window matrix object sketch	Limiting in-memory state and reducing update latency	Streaming (online) analytics and streaming systems

Source: compiled by the authors based on D. Ahfock *et al.* (2023), M. Udell & Z. Frangella (2023), H. Yin *et al.* (2024)

Table 2 summarises three complementary directions of application of approximate LA methods in information systems: RNLA as a way to reduce the resource cost of calculations, theoretical properties of sketching as the basis of reliability, and sliding-window sketching as a mechanism for working in streaming modes. The common basis is the replacement of operations on full-dimensional matrix objects with calculations on compact representations with a controlled error. RNLA is focused on acceleration and savings of CPU/memory/I/O through sketch-representation of matrices and operators in batch analytics problems. Analysis of sketching properties

provides probabilistic quality guarantees (for norms, subspaces and spectral characteristics), which supports parameter selection and control of degradation risk. The sliding-window approach provides incremental sketch updates when the data "window" changes, limiting the state in memory and reducing the latency of updates in streaming systems. The system value of RNLA/sketching is determined by a combination of three components: practical cost reduction, availability of quality guarantees and adaptation to the processing mode (batch or streaming). In short, approximate LA methods are not only mathematical techniques, but also an engineering scaling tool that allows balancing the accuracy of results with memory limitations, I/O and execution time requirements. Along with approximate calculations through RNLA and sketching, a critical direction for optimising resources in systems is the compression of the matrix representations – both data and model parameters. Compression of matrices and models in information systems comes down to two complementary mechanisms: low-rank, when a matrix object is approximated by a lower internal dimension, and low-precision, when the bit depth of the element representation is reduced. Both approaches reduce memory and bandwidth requirements, and can also speed up matrix operations with a controlled loss of accuracy (Table 3).

Table 3. Comparative characteristics of low-rank and low-precision matrix and model compression approaches

Approach	Compression principle	Quality management	Main system-level effect
Weighted low-rank factorisation (for compression of language models)	Low-rank approximation of parametric matrices	Weights in the loss function specify uneven approximation accuracy (component prioritisation)	Reducing the amount of parameters and potentially speeding up multiplications with controlled loss of quality
Randomised low-rank factorisation with quantisation (low precision)	Combination of structural (rank) and numerical (precision) compression	The rank and bitness parameters determine the "accuracy-resources" tradeoff, criticality is the total consideration of errors	Reduced memory footprint and memory traffic + possible acceleration of calculations on low-precision hardware

Source: compiled by the authors based on Y.-C. Hsu *et al.* (2022), R. Saha *et al.* (2023)

Weighted low-rank factorisation is focused on problem-specific quality preservation: the weights in the loss function allow for differentially distributing the approximation error, providing more accurate reproduction of critical components and more "aggressive" compression of less significant ones. Randomised low-rank and low-precision factorisation combines two independent compression mechanisms – structural (rank reduction) and numerical (bit size reduction), which directly reduces the storage volume of factors and traffic between memory levels, and can also increase computational efficiency on hardware optimised for low precision. At the same time, the combination of low-rank and low-precision increases the requirements for total error control, since the contributions of rank approximation and quantisation accumulate. In general, factorisation compression can be constructed as an adaptive precision control (through weights) to preserve quality in the target problem, or as a system-oriented combined compression (low-rank+low-precision) aimed at minimising memory footprint and memory access costs. Both approaches implement the "precision-resources" trade-off, but do so with different "leverages": weighted low-rank factorisation – through component prioritisation, randomised low-rank factorisation with quantisation – through simultaneous limitation of the internal dimensionality and bitness of the representation. This emphasises the role of linear algebra as a tool for engineering optimisation, where factorisation parameters serve as a means of tuning performance to specific system constraints.

The results of the study showed that the effectiveness of LA optimisations in sparse computing is determined not only by the correctness of algebraic rewrites, but also by the execution mode, in particular, the way of materialising intermediate results and organising memory access. This is consistent with the conclusions of A. Mastoras *et al.* (2023), who found that the transition to nonblocking semantics in GraphBLAS makes deferred execution and fusion of sequences of operations possible, which reduces the volume of intermediate structures and changes the memory access profile. The authors also showed that the performance gain is parameter-dependent: it is determined by the configuration of parallelism and block sizes, and can degrade in the event of an unsuccessful choice due to the overhead of execution management. The approach taken in both studies concretises the "bottleneck" between the optimiser and execution: algebraically equivalent plans require execution mechanisms that are able to transform rewrites and planning into a real reduction in memory/I/O pressure and obtain measurable speedup.

The results of the study showed that structural low-rank compression of tensor representations reduces memory and storage requirements at the cost of a controlled loss of accuracy, and the effect is determined by the approximation parameters. This is consistent with the study by T. Kwon *et al.* (2025), who proposed TensorCodec based on neural tensor-train decomposition and showed that the compression quality is further improved by folding into a higher-order tensor and reordering modes to better detect structural dependencies. The authors experimentally demonstrated an improvement in the "compression degree – reconstruction error" trade-off compared to the basic alternatives, which

confirms the parameterised nature of tensor compression: resource savings are achieved not "for free", but through a controlled choice of the compression level, which directly specifies the permissible quality degradation. A sustainable acceleration effect is achieved only with a coordinated optimisation of plans, execution primitives and infrastructure that removes network and I/O constraints.

In the study by Z. Shen *et al.* (2025), the researchers found that the use of matrix operations in storage systems (via erasure coding) is key to overcoming resource constraints while ensuring fault tolerance. The authors emphasised that the evolution of coding methods is aimed at minimising the I/O overhead during data recovery, which requires deep integration of LA primitives into the physical layer of the system. This correlates with the results of this study on the role of linear algebra as an engineering optimisation tool, where the choice of storage format and materialisation strategy determines the I/O profile of the entire system. Both approaches emphasise that a sustainable speedup effect is achieved only with a coordinated optimisation of plans, execution primitives, and infrastructure that removes network and I/O constraints. Thus, linear algebra unifies data representation and analytical/ML procedures in the form of LA expressions, which allows for correct rewrites and cost-effective plan selection with minimal intermediate materialisation. The practical effect is determined by the consistency of rewriting rules, cost model, and physical implementation, which sets real trade-offs between performance, memory/I/O, and accuracy.

Optimisation of storage, access, and execution of LA computations

Matrix representations are natural in big data problems, but matrices are mostly sparse, so dense storage incurs unnecessary memory and I/O overhead and slows down computations by handling zero elements. Sparse formats store only non-zero values and indices, reducing data size and the cost of passes. Critically, the choice of sparse format affects not only compactness but also the cost of LA operations: formats optimise row/column access, updates, and parallelisation differently, determining which operations will be relatively "cheap" or "expensive" in LA pipelines (Marichal *et al.*, 2021). Columnar storage is fundamental to analytical systems, as queries selectively read attributes and perform scans with filters and aggregations, with performance determined by the speed of reading the required columns, the overhead of encoding/decompression, and cache management, i.e., the storage format as a key factor in the I/O profile. For LA workloads, this means that feature preparation and data transformation often boil down to intensive scans, where I/O and decoding can dominate arithmetic. Accordingly, the throughput of reading/processing columns determines the cost of building matrix representations from tabular data (Zeng *et al.*, 2023).

Unlike sparse and columnar, which optimise physical storage at the element or column level, factorised databases reduce redundancy at the result-structure level. In queries with joins and groupings, factorised representations compactly encode dependencies and common subexpressions instead of fully materialising all combinations, which potentially reduces the size of intermediate results, memory pressure, and I/O, provided that the optimiser and execution layer support it (Olteanu, 2023). In the context of LA workloads, factorised representations are a complementary level of compactness alongside sparse (zero minimisation) and columnar (efficient attribute scans): together, these approaches emphasise the multi-level nature of storage optimisation and its impact on the performance of LA-oriented plans (Marichal *et al.*, 2021; Olteanu, 2023). The spread of embeddings has formed a separate class of vector database management systems focused on storing high-dimensional vectors and similarity queries (Nearest Neighbours (NN)/Approximate Nearest Neighbour (ANN)) or hybrid queries with metadata filters. The performance is determined not only by the search algorithms, but also by the storage architecture, indexes, query execution framework and resource management. A typical architecture includes storage subsystems (vector/metadata placement), indexing (ANN and the "accuracy-speed" trade-off), execution (scheduling, filters, similarity calculations) and background support (reindexing, compaction), with the computational foundation directly relying on linear algebra operations (Pan *et al.*, 2024).

In practical vector search, the bottleneck is not similarity calculations, but I/O and the cost of data access. In storage-based ANNs, vectors and indices do not fit entirely in RAM, so performance is determined by storage characteristics, access patterns, and caching. Under such conditions, ANN should be treated as a system optimisation problem that requires coordinated co-design of indices, physical representation, compression/quantisation, and execution strategies; local speedup of one component may not yield any benefit without coordination with others. Thus, vector search is a class of workloads that combine LA computations, compression, and access optimisation, and the result is determined by the system consistency of these components (Pan *et al.*, 2023; Pan *et al.*, 2024). The effectiveness of LA methods depends not only on the optimiser, but also on the execution primitives. Standardised programming interfaces specify a unified set of operations and create a basis for low-level optimisations (parallelisation, vectorisation, and sparsity). GraphBLAS formalises matrix/vector operations and parameterises semantics via semirings, acting as a "bridge" between LA representation and productive execution; standardisation also increases portability of optimisations across implementations (Brock *et al.*, 2021).

In distributed systems, performance is limited not by arithmetic but by shuffling data between nodes, so

LA optimisations can lose the effectiveness if infrastructure costs dominate. DShuffle is an example of an infrastructure optimisation of data redistribution in large-scale processing, where the bottleneck is the internode exchange of intermediate data and the overhead of the central processor. The approach is based on the use of a data processing unit (DPU) to offload some operations, which reduces the load on the CPU and improves system performance (traffic, throughput) in distributed scenarios (Ding *et al.*, 2025). Together, GraphBLAS and DPU-oriented shuffle illustrate the principle of co-design: sustained speedup is achieved by concerted optimisation of LA plans, execution primitives, and infrastructure that removes network and I/O constraints. Based on the approaches considered, a generalisation of linear algebra approaches is presented in Table 4 by application areas, types of resource constraints, metrics, and risks.

Table 4. Generalisation of linear algebra approaches for optimising data processing and storage in modern 21st century information systems

Direction of application	Resource constraint type	Method (linear algebra/related)	Principle of operation	Target metrics	Limitations and risks
Optimisation of analytical/ML query execution in database management systems	Materialisation of intermediate results, ineffective plans	Linear algebraic query processing, expression rewriting, dynamic scheduling	Equivalent expression transformations, operator fusion; intermediate data reduction	Response time, bandwidth, read/write (I/O), RAM	Cost model, integration complexity
Approximate matrix methods for accelerating calculations	CPU, sometimes RAM	RNLA, randomised sketching	Random projections/sampling, dimensionality reduction	Computation time, RAM, error	Accuracy-speed tradeoff, parameter selection, dependence on data properties
Streaming processing (sliding window)	CPU for calculation, RAM	Sketching matrices on a sliding window	Incremental sketch updates	Response time, bandwidth; RAM, error	Error control, dependence on window parameters, complexity of implementation
Matrix/Model Compression	RAM/video memory, storage, memory bandwidth	Low-rank (weighted) factorisation, low-rank+low-precision factorisation	Approximation with fewer parameters, quantisation	Size, memory usage, speed, error	Quality deterioration, accumulation of errors, validation, and calibration required
Physical representation of big data	I/O, RAM	Sparse storage (compressed rows/columns), columnar formats	Storing non-zeros, reading the required columns, compression	I/O, volume, query time	Expensive upgrades, possible CPU bottleneck shift (decompression)
Vector databases and similarity search	Mostly I/O, indexes, storage	Vector database management systems, I/O-aware ANNs	Indexing; reading reduction, candidate selection	Response time, I/O, volume, accuracy/completeness	Accuracy-speed tradeoff, vector distribution dependency, I/O dominance

Notes: Materialisation (materialisation of intermediate results) – execution of a query operator with physical formation and temporary storage of its output data set (e.g., in a temporary table or file) for further processing by subsequent operators
Source: compiled by the authors based on R. Marichal *et al.* (2021), Y.-C. Hsu *et al.* (2022), R. Saha *et al.* (2023), D. Ahfock *et al.* (2023), M. Udell & Z. Frangella (2023), X. Zeng *et al.* (2023), B. Ding *et al.* (2024), H. Yin *et al.* (2024), J.J. Pan *et al.* (2024), W. Sun *et al.* (2025), T. Muñoz (2025), Z. Ren *et al.* (2025)

LA-oriented optimisation in information systems is reduced to managing the interrelated trade-offs between performance and correctness of the result. Approximate approaches (RNLA, sketching, low-rank, low-precision) provide a reduction in time and memory costs, but require controlled parameterisation and quality validation procedures, especially in incremental update modes. At the same time, in practical workloads, the determining factors are not arithmetic operations, but the costs of data access and movement (I/O, caches, exchange between nodes), which is manifested in storage-based vector search configurations. Therefore, a stable optimisation effect is achieved only by coordinating optimisations at the expression/plan level with the cost model and physical implementations of operators, otherwise the optimisation remains partial and vulnerable to system bottlenecks.

Given the set of results considered, the choice of LA methods and system mechanisms for the implementation should be made taking into account the dominant resource constraints and the type of workload. For ML queries and data science pipelines, the algebraic representation of calculations in the form of LA expressions is a priority, since it supports rewriting and cost-based selection of plans, reducing the materialisation of intermediate results and integration overhead. Under the conditions of operator evolution and the emergence of domain-specific rules, the extensibility of the optimiser (rewriting rules, cost model, physical implementations) is decisive, without which the optimisation effect becomes unstable. When the execution profile is compute-bound and approximation is permissible, RNLA and matrix sketching are applicable, and in streaming modes, sliding-window sketch maintenance mechanisms are used to maintain quality during incremental updates. When memory and bandwidth constraints dominate, low-rank and low-precision factorisations are appropriate, which reduce the footprint and potentially speed up multiplication while controlling the total error. In terms of storage, the format should be consistent with the operations: sparse and columnar representations determine the I/O profile and decoding overhead, and factorised approaches reduce the structural redundancy of materialisations. For vector search, the co-design of indexing, execution, and storage is critical, particularly in storage-based modes, where the scaling limit is set by I/O characteristics. In general, the effectiveness of LA optimisations depends on the "vertical" consistency of planning, execution primitives (in particular GraphBLAS), and, in distributed scenarios, the infrastructure costs of the network and shuffle.

The results of the study confirmed that in practical vector search workloads, the main bottleneck is I/O costs and data access costs, which directly determine the scaling limits for storage-based systems. The revealed pattern justifies the need for coordinated co-design of indexes, physical representation, and execution strategies as a key means of overcoming infrastructure limitations. This approach is consistent with the findings of Y. Pan *et al*. (2023), who showed that to ensure scalability on large data volumes, the transition to disk-oriented structures with low RAM usage (LM-DiskANN) is critical. The authors argued that optimising graph indexes directly for disk-specificity allows maintaining high performance of ANN queries even under strict RAM limitations. Within the framework of system logic, both works confirm that the real efficiency of vector systems is determined not only by the chosen algorithm, but also by the architectural integrity with the storage, where it is I/O characteristics that become the determining factor of global performance.

X. Bao *et al*. (2024) found that integrating machine learning models directly into the logical rules of a data cleansing system (Rock) reduces the overhead associated with the gap between declarative semantics and procedural ML components. The authors showed that "embedded" execution of ML models within the logical plan avoids external calls, intermediate materialisation, and data duplication, which directly affects performance and predictability of execution. These results are consistent with current research on the need for an algebraic representation of ML calculations as part of the optimisation core of the system, rather than as external black boxes. In particular, the authors' approach illustrates that the extensibility of the optimiser through the introduction of new rules and operators is a critical prerequisite for effective support of LA workloads in data science, since it is at the level of algebraic representation that consistent rewriting, planning, and cost-effective execution selection become possible. At the same time, the approach of X. Bao *et al*. has an applied focus on the data cleaning problem and does not consider the general space of linear algebraic rewrites and cost-based optimisation for a broader class of ML and analytical workloads. In contrast to the system framework formulated in this study, the authors' work is limited to analysing the impact of physical operators and execution parameters (memory, parallelism, I/O) on the global cost of the plan. Integrating ML into the declarative core is feasible, but requires LA-oriented optimisation at the logical, cost, and physical levels.

B. Tian *et al*. (2024) found that offloading part of the ANN search computations to SmartSSDs (solid-state storage) allows for scaling queries to billions of volumes while reducing latency and increasing energy efficiency. The authors showed that in storage-based ANNs, the bottleneck is not the similarity calculation itself, but the data transfer between the storage and the host, so near-storage computing reduces the pressure on the transmission bus and reduces host overhead. The effectiveness of this approach depends on the redistribution of functions between the device and the host (organisation of accesses, candidate selection, execution of some operators on the SSD side), i.e., on the architectural consistency of the execution loop. This correlates with the results of this study on infrastructure optimisation (DPU/DShuffle) as a way to remove network and I/O constraints in distributed scenarios: the acceleration is achieved not by local optimisation of a single operator, but by moving critical processing stages closer to the data and reducing the amount of inter-component traffic. Thus, the obtained results confirm the principle of co-design: a sustainable performance gain is formed under the condition of coordination of the execution plan (including LA operators that implement selection/aggregation/similarity calculation) with the available hardware primitives and the limitations on bandwidth and data exchange.

The results of the study showed that the key factor in the effectiveness of LA optimisations is the "vertical" consistency between the planning process and specific physical execution strategies, especially in conditions of a shortage of fast memory. Minimising the materialisation

of intermediate results by choosing optimal execution plans allows increasing system performance and reducing resource load. This conclusion is supported by the results of the work of Z. Zhang *et al.* (2024), where it was demonstrated that the implementation of ordered pipeline processing allows efficiently executing vector queries, even if the amount of data exceeds the available video memory of the GPU. The authors confirmed that careful control of the order of operations in combination with overlapping calculations by data transfer allows successfully levelling out materialisation delays. The acceleration effect discovered by Z. Zhang *et al.* serves as a practical validation of the approach proposed in this study to the systematic minimisation of intermediate data through the consistency of the planning and execution levels. Thus, the effectiveness of LA-oriented optimisation is determined by the "vertical" consistency of planning with physical execution and the ability to minimise the materialisation of intermediate results; a sustainable effect is achieved only with the operation of a consistent chain "rewriting→cost model→physical execution".

CONCLUSIONS

The results showed that matrix representations are natural for large datasets, but the prevailing sparsity makes dense storage a source of excessive memory and I/O overhead and slowdown due to processing of zero elements. Sparse storage formats reduce the size of the data by storing only non-zero values and indices, reducing the cost of passes. The choice of sparse format affects not only compactness, but also the "price" of LA operations due to different efficiencies of row/column accesses, updates, and parallelisation. For analytical systems, columnar storage determines the I/O profile of queries, and performance depends on the speed of reading the required columns, the overhead of encoding/decompression, and the behaviour of caches. In LA workloads, feature preparation and transformations often boil down to intensive scans, where I/O and decoding can dominate arithmetic. Factorised databases reduce redundancy at the result-structure level by reducing the amount of intermediate materialisations in joins and groupings, provided that such representations are supported by the optimiser and execution layer.

The spread of embeddings has formed a class of vector database management systems in which efficiency is determined not only by the ANN algorithm, but also by the storage architecture, indexes, execution path, and resource management. In storage-based ANNs, the bottleneck is usually not the similarity computation, but I/O and the cost of accessing data when vectors and indexes do not fit entirely in RAM. I/O dominance requires a coordinated co-design of indexes, physical representation, compression/quantisation, and execution strategies, and local speedups of one component may not yield any benefit without coordination with the others. In distributed scenarios, the effect of LA optimisations is limited by the costs of data movement (shuffle), so sustainable acceleration requires simultaneous coordination of LA plans with execution primitives and infrastructure mechanisms for removing network and I/O constraints. The limitation of the study is its theoretical nature, without experimental validation on a single prototype of the information system. A promising direction for future research is the development and experimental validation of an LA-oriented optimisation loop (rewriting + cost model + physical operators), coordinated with storage-based ANN and hardware primitives (DPU/near-storage/SmartSSD) and supplemented by compression and RNLA/sketching methods with formal quality guarantees and automatic parameter selection.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Ahfock, D., Astle, W.J., & Richardson, S. (2023). On randomized sketching algorithms and the Tracy-Widom law. *Statistics and Computing*, 33, article number 34. doi: 10.1007/s11222-022-10148-5.

[2] Bao, X., *et al.* (2024). Rock: Cleaning data by embedding ML in logic rules. In *Companion of the 2024 international conference on management of data* (pp. 106-119). New York: Association for Computing Machinery. doi: 10.1145/3626246.3653372.

[3] Boehm, M., Interlandi, M., & Jermaine, C. (2023). Optimizing tensor computations: From applications to compilation and runtime techniques. In *Companion of the 2023 international conference on management of data* (pp. 53-59). New York: Association for Computing Machinery. doi: 10.1145/3555041.358940.

[4] Brock, B., Buluç, A., Mattson, T.G., McMillan, S., & Moreira, J. (2021). *The GraphBLAS C API specification: Version 2.0.0*. Retrieved from https://graphblas.org/docs/GraphBLAS_API_C_v2.0.0.pdf.

[5] Ding, B., Narasayya, V., & Chaudhuri, S. (2024). Extensible query optimizers in practice. *Foundations and Trends in Databases*, 14(3-4), 186-402. doi: 10.1561/1900000077.

[6] Ding, C., Li, S., Lu, K., Yao, T., Wang, D., Wu, H., Wan, J., Tan, Z., & Xie, C. (2025). DShuffle: DPU-optimized shuffle framework for large-scale data processing. In *2025 USENIX annual technical conference* (pp. 1371-1386). Boston: USENIX Association.

[7] Geerts, F., Muñoz, T., Riveros, C., & Vrgoč, D. (2021). Expressive power of linear algebra query languages. In *Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI symposium on principles of database systems* (pp. 342-354). New York: Association for Computing Machinery. doi: 10.1145/3452021.3458314.
[8] GraphBLAS Forum. (n.d.). Retrieved from https://graphblas.org/.
[9] Hordiichuk-Bublivska, O.V., & Fabri, L.P. (2022). Matrix factorization of big data in the industrial systems. *Ukrainian Journal of Information Technology*, 4(2), 68-73. doi: 10.23939/ujit2022.02.068.
[10] Hsu, Y.-C., Hua, T., Chang, S.-E., Lou, Q., Shen, Y., & Jin, H. (2022). Language model compression with weighted low-rank factorization. *ArXiv*. doi: 10.48550/arXiv.2207.00112.
[11] Jin, G., Bian, H., Chen, Y., & Du, X. (2022). Columnar storage optimization and caching for data lakes. In *Proceedings of the 25th international conference on extending database technology* (pp. 419-423). Edinburgh: Open Proceedings. doi: 10.48786/edbt.2022.33.
[12] Kwon, T., Ko, J., Jung, J., Jang, J.-G., & Shin, K. (2025). Compact lossy compression of tensors via neural tensor-train decomposition. *Knowledge and Information Systems*, 67, 1169-1211. doi: 10.1007/s10115-024-02252-x.
[13] Lane, P.A., & Booth, J.D. (2023). Heterogeneous sparse matrix-vector multiplication via compressed sparse row format. *Parallel Computing*, 115, article number 102997. doi: 10.1016/j.parco.2023.102997.
[14] Li, Z., Sun, W., Zhan, D., Kang, Y., Chen, L., Bozzon, A., & Hai, R. (2024). Amalur: The convergence of data integration and machine learning. *IEEE Transactions on Knowledge and Data Engineering*, 36(12), 7353-7367. doi: 10.1109/TKDE.2024.3357389.
[15] Liu, C., Pavlenko, A., Interlandi, M., & Haynes, B. (2025). Data formats in analytical DBMSs: Performance trade-offs and future directions. *The VLDB Journal*, 34, article number 30. doi: 10.1007/s00778-025-00911-1.
[16] Marichal, R., Dufrechou, E., & Ezzatti, P. (2021). Optimizing sparse matrix storage for the big data era. In M. Naiouf, E. Rucci, F. Chichizola & L. De Giusti (Eds.), *9th conference: Cloud computing, big data & emerging topics* (pp. 121-135). Cham: Springer. doi: 10.1007/978-3-030-84825-5_9.
[17] Mastoras, A., Anagnostidis, S., & Yzelman, A.-J.N. (2023). Design and implementation for nonblocking execution in GraphBLAS: Tradeoffs and performance. *ACM Transactions on Architecture and Code Optimization*, 20(1), article number 6. doi: 10.1145/3561652.
[18] Muñoz, T. (2025). Enumeration-based dynamic query processing for linear algebra workloads in data science. In *VLDB 2025: PhD workshop*. London: QEII Centre.
[19] Olteanu, D. (2023). Report on the workshop on factorized databases. *SIGMOD Record*, 52(2), 53-56.
[20] Pan, J.J., Wang, J., & Li, G. (2024). Survey of vector database management systems. *The VLDB Journal*, 33, 1591-1615. doi: 10.1007/s00778-024-00864-x.
[21] Pan, Y., Sun, J., & Yu, H. (2023). LM-DiskANN: Low memory footprint in disk-native dynamic graph-based ANN indexing. In *2023 IEEE international conference on big data* (pp. 5987-5996). Sorrento: IEEE. doi: 10.1109/BigData59044.2023.10386517.
[22] Ren, Z., Doekemeijer, K., Apparao, P., & Trivedi, A. (2025). Storage-based approximate nearest neighbor search: What are the performance, cost, and I/O characteristics? In *2025 IEEE international symposium on workload characterization* (pp. 407-422). Irvine: IEEE. doi: 10.1109/IISWC66894.2025.00041.
[23] Saha, R., Srivastava, V., & Pilanci, M. (2023). Matrix compression via randomized low rank and low precision factorization. In *Proceedings of the 37th international conference on neural information processing systems* (pp. 18828-18872). Red Hook: Curran Associates Inc.
[24] Serrano, T.M., Riveros, C., & Vansummeren, S. (2024). Enumeration and updates for conjunctive linear algebra queries through expressibility. In *Proceedings of the 27th international conference on database theory (ICDT 2024)*. Wadern: Dagstuhl Publishing. doi: 10.4230/LIPIcs.ICDT.2024.12.
[25] Shen, Z., Cai, Y., Cheng, K., Lee, P.P.C., Li, X., Hu, Y., & Shu, J. (2025). A survey of the past, present, and future of erasure coding for storage systems. *ACM Transactions on Storage*, 21(1), article number 4. doi: 10.1145/3708994.
[26] Striamets, S.P., & Opotyak, Yu.V. (2021). Matrix parallel processor based on a homogeneous computational medium using an advanced computing cell. *Ukrainian Journal of Information Technology*, 3(1), 78-84. doi: 10.23939/ujit2021.03.078.
[27] Sun, W., Katsifodimos, A., & Hai, R. (2025). Accelerating machine learning queries with linear algebra query processing. *Distributed and Parallel Databases*, 43, article number 8. doi: 10.1007/s10619-024-07451-7.
[28] Tian, B., Liu, H., Duan, Z., Liao, X., Jin, H., & Zhang, Y. (2024). Scalable billion-point approximate nearest neighbor search using SmartSSDs. In *Proceedings of the 2024 USENIX annual technical conference* (pp. 1135-1150). San Francisco: USENIX Association.
[29] Udell, M., & Frangella, Z. (2023). *Randomized numerical linear algebra for optimization*. Retrieved from https://web.stanford.edu/~udell/doc/udell2022_randnla4opt.pdf.

[30] Yin, H., Wen, D., Li, J., Wei, Z., Zhang, X., Huang, Z., & Li, F. (2024). Optimal matrix sketching over sliding Windows. *Proceedings of the VLDB Endowment*, 17(9), 2149-2161. doi: 10.14778/3665844.3665847.
[31] Zeng, X., Hui, Y., Shen, J., Pavlo, A., McKinney, W., & Zhang, H. (2023). An empirical evaluation of columnar storage formats. *Proceedings of the VLDB Endowment*, 17(2), 148-161. doi: 10.14778/3626292.3626298.
[32] Zhang, Z., Liu, F., Huang, G., Liu, X., & Jin, X. (2024). Fast vector query processing for large datasets beyond GPU memory with reordered pipelining. In *Proceedings of the 21st USENIX symposium on networked systems design and implementation* (pp. 23-40). Santa Clara: USENIX Association.

Застосування методів лінійної алгебри для оптимізації обробки та зберігання даних у сучасних інформаційних системах

Олександр Крайчук
Кандидат фізико-математичних наук, професор
Рівненський державний гуманітарний університет
33028, вул. Степана Бандери, 12, м. Рівне, Україна
https://orcid.org/0000-0002-5987-0460
Сергій Крайчук
Кандидат технічних наук, доцент
Рівненський державний гуманітарний університет
33028, вул. Степана Бандери, 12, м. Рівне, Україна
https://orcid.org/0000-0001-9756-1979
Наталія Остапчук
Кандидат педагогічних наук, доцент
Рівненський державний гуманітарний університет
33028, вул. Степана Бандери, 12, м. Рівне, Україна
https://orcid.org/0000-0001-8827-8741

Анотація. Метою дослідження було узагальнити підходи до узгодження рішень на етапах планування, виконання та фізичного рівня подання даних в інформаційних системах за домінування ресурсних обмежень під час оперування масивними наборами. Методологія базувалася на системно-аналітичній формалізації, порівняльно-аналітичному узагальненні та типологізації, порівнянному аналізі, систематизації та інтегративному синтезі. Встановлено, що лінійна алгебра формалізує подання даних як матриць/векторів і обчислення як послідовність перетворень. Узагальнення операцій (напівкільця) поширює ці примітиви на графові та булеві задачі, а Graph Basic Linear Algebra Subprograms стандартизує їх виконання для розріджених структур. Ефективність визначається узгодженим ланцюгом "переписування → вартісна модель → фізичне" виконання: переписування зменшує проміжні дані, а вартісна модель обирає план за введення/виведення даних, пам'яттю, розрідженістю та паралелізмом, без неї лишаються евристики. Виявлено, що факторизовані подання зменшують надлишковість і тиск на пам'ять та введення/виведення даних (для приєднань/групувань) за підтримки системою. Randomized Numerical Linear Algebra і скетчинг знижують витрати та керують компромісом "точність-ресурси" в потоковій обробці. Компресія матриць (низькорангове подання, знижена точність) зменшує пам'ять і трафік та прискорює обчислення за контрольованої втрати точності. У векторних системах керування базами даних і розподілених системах вузьким місцем є введення/виведення даних і перерозподіл між вузлами, тому DShuffle переносить частину обробки ближче до даних, підкреслюючи роль інфраструктурної оптимізації. Лінійно-алгебраїчне подання обчислень створює основу для узгодженої оптимізації переписування, вибір планів і виконання, зменшуючи матеріалізацію та переміщення даних за обмежень введення/виведення, пам'яті й пропускної здатності. Практична значимість результатів полягає у можливості їх застосування розробниками та інженерами інформаційних систем для узгодженого вибору форматів даних, планів виконання й обчислювальних примітивів, щоб зменшити матеріалізацію та витрати на введення/виведення, пам'ять і мережевий обмін

Ключові слова: операції; матриці; ресурсні обмеження; розрідженність; матеріалізація; бази даних

BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 2. 2026

UDC 004.31:621.3:539.1:004.94
DOI: 10.62660/bcstu/2.2026.97

ISSN: 2306-4412
E-ISSN: 2708-6070

Physical aspects of the IT industry: From software architectures to cyber-physical systems

Taras Liashuk*

PhD in Physics and Mathematics, Senior Lecturer
Rivne State University of Humanities
33028, 12 Stepana Bandery Str., Rivne, Ukraine
https://orcid.org/0000-0002-2242-7537

Abstract. The functioning of modern computing systems is based on a complex interconnection of hardware and applied physical phenomena. The relevance of the study was determined by the fundamental dependence of technological innovations in the IT industry on the laws of physics, which simultaneously defined both limits (e.g. quantum tunnelling in nanometre-scale transistors) and new opportunities (e.g. quantum superposition in qubits). The aim of the work was to analyse the interconnection of fundamental physical laws with the hardware-software architecture of computing systems and to determine directions for future development. The methodology of the study was based on a systematic approach, including content analysis of high-quality scientific data, structural-functional analysis of physical clusters, and theoretical synthesis of interdisciplinary connections, supported by logical-graphical modelling. The main results of the study revealed that quantum mechanics and thermodynamics set the physical limits for transistor miniaturisation and managed the thermal regime of microchips, while optics was the foundation for high-speed data transmission and microchip lithography. Classical mechanics and electrodynamics ensured the functioning of cyber-physical systems, including automated control systems, robotic complexes, and wireless communication infrastructure. It was analysed that molecular physics and the modelling of nanoscale structures played a crucial role in the design of new semiconductor materials and the prediction of component durability based on appropriate computer-physical models. It was shown that nuclear and atomic physics played a key role in ensuring the radiation hardness of electronics and were considered a potential power source for highly autonomous computing complexes. Particular attention was paid to promising areas, specifically superconductivity, quantum computing, and quantum cryptography. The practical value of the work lies in forming a comprehensive perspective on the physical foundations of computing systems, which was critically important for computer and software engineers, and scientists when developing new materials, designing high-performance and energy-efficient microchips, and creating reliable autonomous and quantum computing systems

Keywords: computational physics; nanotechnology; computer modelling; planar technology; quantum computing; energy efficiency; signal integrity

INTRODUCTION

Global technological progress is characterised by the rapid approach of information and computing systems (ICS) to critical physical limits, demanding an integrated approach to modelling and practical implementation of materials, thermodynamics, and high-frequency electrodynamics at the nanoscale. The necessity for further transistor miniaturisation, enhanced energy efficiency, and the development of novel software architectures requires a deep understanding of subatomic processes, where computer modelling serves as the

Article's History: Received: 01.12.2025; Revised: 30.03.2026; Accepted: 18.05.2026; Published: 26.06.2026

Suggested Citation:
Liashuk, T. (2026). Physical aspects of the IT industry: From software architectures to cyber-physical systems. *Bulletin of Cherkasy State Technological University*, 31(2), 97-108. doi: 10.62660/bcstu/2.2026.97.

*Corresponding author(taras.liashuk@rshu.edu.ua)

Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/)

foundation for industrial prototyping, optimising device behaviour, and ensuring the structural integrity of final hardware-software solutions. Given this, the relevance of the study lies in the necessity to systematise the fundamental physical principles underlying the development of innovative information technology (IT) solutions, including enhancing thermal efficiency, ensuring high-speed signal integrity, and developing post-silicon computing paradigms. The objective of the work was to conduct comprehensive analysis of the influence of various fields of computational physics (CP) on the formation of ICS architecture.

One of the most acute problems is overcoming the physical limits of transistor miniaturisation. In the work of Y. Cai *et al*. (2020), it was shown that the study of nanoscale semiconductors (SCs) required the application of the laws of quantum mechanics (QM) to predict electron behaviour. In the works of P. Kumari (2024) and A. Mubashr (2025), dedicated to the physical limits of miniaturisation and quantum tunnelling, the necessity of transitioning to new materials, whose properties were modelled using Ab Initio methods (quantum chemistry), was emphasised. The issue of thermal regime control for high-performance chips, which was a direct application of thermodynamics, was actively investigated in the context of two-phase cooling in the work of N. Mohammed *et al*. (2025). The analysis demonstrated that, in liquid cooling, the maximum server temperature diminished to 65°C, and power usage effectiveness improved from 2 in the air cooling system to 1.01 in liquid cooling. The application of copper coating resulted in a 13% enhancement in the gradual temperature distribution among servers, especially in high-power servers, and eliminated hottest spots by decreasing the maximum temperature. However, the simulation indicated a comparatively modest temperature improvement of 8% when coating low-power servers. Continuing the topic, the reliability assessment of systems under thermal loads was also considered by H.-S. Kim *et al*. (2025). It was found that the fatigue life and fatigue crack type (solder joint crack or Cu trace crack) were influenced and changed significantly by package geometry and solder joint materials. The predicted fatigue life and crack type from Finite Element Analysis (FEA) agreed well with the experimental results.

In the field of communications, the study by D. Kalavathi (2025) was focused on optimising signal integrity and electromagnetic compatibility through numerical modelling of electrodynamics for high frequency ranges. A comprehensive approach to modelling and optimising electromagnetic interference was developed using deterministic and statistical methods alongside 3D electromagnetic solvers. Advanced optimisation techniques, including Genetic Algorithms and Machine Learning-based predictive models, were employed to minimise electromagnetic interference while ensuring system performance. The effectiveness of the proposed design methodologies was confirmed through experimental validation using hardware prototypes and electromagnetic interference test setups. Further development of optoelectronics and LiDAR (Light Detection and Ranging) systems was demonstrated in the study by P. Pendse (2024), which was based on the improvement of optics and sensors, being critically important for autonomous vehicles.

The most ambitious areas, such as quantum computers based on superconducting qubits, were a direct synthesis of QM and superconductivity, highlighted in the work O. Ezratty (2023). A great variety of technological options was demonstrated, explaining how superconducting qubit chipsets were manufactured, describing challenges related to qubit control via classical electronics, and analysing ways to improve their accuracy and optimise their energy footprint. Also relevant were studies of the impact of cosmic radiation (nuclear physics) on the durability and reliability of electronics, which had direct application in space technology and was shown by the authors F.N.U. Naveed & J. Dix (2025). This paper presented a radiation-hardened 4-bit flash analogue-to-digital converter implemented in a 22 nm fully depleted silicon-on-insulator process for high-reliability applications in radiation environments. The analysis of scientific and technical literature confirmed the decisive role of the interconnectedness of physics and IT, demonstrating that fundamental physical limitations and capabilities determined the vectors of development for ICS hardware and software architecture. Recent studies in the most advanced, yet most problematic, areas of CP were reviewed. The research problematics were focused on the fact that traditional engineering design approaches increasingly encountered quantum and relativistic effects (Hidalgo, 2022), which were previously ignored. This necessitated the integration of numerical methods for accurate modelling (Camiola *et al.*, 2025).

The scientific novelty of the work consisted in the systematisation and integrative representation of the role of fundamental physical disciplines, spanning the range from classical to quantum physics, to create a unified, interdisciplinary concept for the functioning and development of key technological areas within the IT industry. Given the interdisciplinary nature of the identified scientific problems and the lack of their comprehensive unification, the aim of the work was to generalise and integrate knowledge regarding the key physical phenomena underlying the architecture, materials science, reliability, and functional capacity of modern ICS.

MATERIALS AND METHODS

The methodology of this research was based on a comprehensive interdisciplinary approach aimed at the systematisation and integration of fundamental physical laws within the context of the development and functioning of modern ICS. To achieve the set objective, a multi-stage research strategy was employed, which

included the selection, critical analysis, and synthesis of high-quality scientific data at the intersection of CP and IT. At the first stage, a systematic literature search and data collection process was carried out. The information base of the study was formed using reputable international scientometric databases, including Scopus, Web of Science, IEEE Xplore, Google Scholar, and others. The data collection was specifically focused on journals and monographs specialising in IT and physics, particularly those exploring their deep interconnection. The search depth was primarily focused on the period between 2020 and 2025 to ensure the relevance of the data regarding the latest technological paradigms, such as 22 nm fully depleted silicon-on-insulator processes, extreme ultraviolet lithography (EUV), and superconducting quantum computing. The selection of sources was performed based on keyword queries, including "quantum effects in semiconductors", "thermal management in high-performance computing", "signal integrity in gigahertz ranges", "post-silicon computing paradigms", and others. At the second stage, the method of content analysis and critical evaluation was applied to the selected scientific materials. Each source was analysed for its contribution to the understanding of how specific physical phenomena (e.g., quantum tunnelling, thermal conductivity, or electromagnetic interference) influence the architecture and reliability of hardware and software solutions. Special attention was paid to identifying "bottlenecks" where traditional engineering approaches encountered fundamental physical limits, such as the diffraction limit in optics or the sub-threshold swing in transistors.

The third stage involved the use of a systemic approach and structural-functional analysis to create a unified interdisciplinary concept. Fundamental physical disciplines were categorised into six key clusters: classical and relativistic mechanics, molecular physics and thermodynamics, electricity and magnetism, optics, QM, and atomic and nuclear physics. For each cluster, the direct application in IT was identified, ranging from the mechanical reliability of server enclosures to the quantum-mechanical modelling of qubits. This allowed for the transition from a fragmented description of physical effects to an integrated model of ICS functioning as a complex physical-technical system. To ensure the clarity of the proposed conceptual framework, a method of logical-graphical modelling was employed. This involved the structural mapping of causal relationships between physical laws and IT hardware parameters, resulting in the creation of integrated diagrams that synthesise complex interdisciplinary data into a coherent visual format.

At the final stage, the method of theoretical synthesis and generalisation was utilised to formulate the conclusions of the study. Comparative analysis was used to evaluate different modelling methods; specifically, FEA was identified as the most effective for large-scale mechanical and thermal simulations due to its computational efficiency, whereas Density Functional Theory (DFT) provided superior accuracy for predicting the atomic properties of new SC materials (e.g., 2D Janus LiMSSe), despite its significantly higher computational resource requirements. The integration of the obtained data was carried out to define the most promising vectors for the future development of the IT industry, including quantum cryptography and energy-efficient superconducting electronics. Thus, the applied complex of theoretical and analytical methods ensured the validity and comprehensiveness of the formed interdisciplinary vision.

RESULTS AND DISCUSSION

The obtained results presented a systematised analysis of the fundamental and applied areas of CP, detailed the discussion of their foundational principles and specific examples of application in the design, functioning, and development of modern ICS and their hardware and software architecture. Classical mechanics, based on Newton's laws, constitutes a vital foundation for CP by providing the mathematical apparatus for the numerical modelling of the dynamics and evolution of physical systems over time. Through the numerical integration of differential equations of motion, ICS can accurately reproduce the complex behaviour of objects ranging from nanoparticles to astronomical macro-bodies. In the IT industry, mechanics is essential for hardware reliability and real-world simulation, powering physical engines in video games and ensuring precision in high-tech manufacturing, such as the design of moving parts in hard drives or robotic systems.

One of the primary areas of application is hardware design, which focuses on the laws of dynamics, statics, and the strength of materials to ensure the reliability and durability of physical components. The mechanical reliability of enclosures and printed circuit boards (PCBs) is achieved through the principles of statics, where CP utilises the Finite Element Method (FEM) to analyse stresses and deformations. This allows for modelling how electronic boards and fastenings react to external forces, such as impacts or drops, as well as internal loads caused by temperature fluctuations. Furthermore, engineers had to model vibrational resistance to identify and avoid dangerous resonant frequencies that could lead to the failure of solder joints, a factor particularly critical for server equipment and mobile transport devices (Gharaibeh & Gharaibeh, 2023). The dynamics of high-speed moving parts in IT devices also require the direct application of Newtonian laws and kinematics. In Hard Disk Drives (HDDs), CP models the dynamics of the read/write head, which must be positioned over a magnetic platter at distances of less than 1 μm at high speeds; such simulations helped minimise vibrations and deviations that could could data errors (Chen *et al.*, 2025). Similarly, mechanics is used in optical drives (CD/DVD/Blu-ray) to model lens and laser mechanism movements for precise focusing, as well as

in printers and plotters to ensure the reliability of paper feed systems and electric motors.

The principles of mechanics are further extended to Microelectromechanical Systems (MEMS), which underpin the operation of smartphones, unmanned aerial vehicles (UAVs), and navigation systems. Sensors such as accelerometers and gyroscopes measure acceleration and angular velocity using the principle of inertia from Newton's first law. In these devices, CP modelled the movement of microscopic mechanical masses that changed their physical configuration under external forces, which resulted in measurable changes in quantities such as electrical capacitance (Kempe, 2011). Molecular and N-body dynamics also rely on numerical methods within CP, utilising Newton's second law to model systems consisting of a large number of interacting objects. In molecular dynamics, the motion and interaction of atoms and molecules are calculated using intermolecular potentials, enabled the study of material properties or protein folding (Pal & Ray, 2020). Conversely, N-body dynamics is applied to larger systems where interactions are described by gravitational or electrostatic forces, which allowed researchers to predict the behaviour of systems, such as electron beam collisions in plasma, that are nearly impossible to investigate through physical experiments (Patsis & Okalidis, 2025).

In the field of robotics and automated systems, classical mechanics serves as the basis for control algorithms that ensure precise task execution. This involves kinematic calculations to describe manipulator motion without considering forces, as well as dynamic analysis that accounts for gravity, friction, and inertia. Computing systems had to numerically determine the torque and power required for each motor to achieve smooth and energy-efficient movement (Kadri *et al.*, 2025). Furthermore, virtual simulation allowed for the creation of precise digital twins of robots to plan trajectories and model collisions using the laws of momentum and energy conservation (Wang *et al.*, 2022). For mobile systems such as UAVs, mechanics is constantly used to calculate the centre of mass and stability, which allowed control systems to instantaneously compensate for external factors like wind or tilt, ensuring balance and preventing loss of stability (Yunping *et al.*, 2018). Realism in computer graphics and game engines is achieved through specialised physical engines that utilise mechanical principles for real-time behaviour modelling. These engines constantly calculate forces such as gravity, friction, and air resistance, applying conservation laws to model complex interactions like object collisions, explosions, and fluid dynamics (Sung *et al.*, 2025).

Finally, while most engineering solutions rely on classical mechanics, certain IT systems involving space and high speeds must account for Einstein's theory of relativity. Global Navigation Satellite Systems (GNSS), such as Global Positioning System (GPS), require constant signal adjustments based on relativistic effects. Special relativity accounts for the high speed of satellites (≈ 14,000 km/h) causing onboard clocks to run slower, while general relativity accounts for the weaker gravitational field causing them to run faster. If ignored, these effects would lead to positioning errors exceeding 10 km per day; therefore, computational components numerically calculated these relativistic corrections to maintain the precision required for modern navigation (Ayso & Kahveci, 2025).

Molecular physics and thermodynamics constitute an indispensable foundation for CP, as they govern all energetic, thermal, and structural processes in IT equipment at both micro- and macroscopic levels. These branches of physics provide the mathematical apparatus, based on statistical laws and the principles of energy conservation, for modelling the behaviour of atoms and molecules. Such modelling is critically important for understanding the thermal balance, efficiency, and longevity of systems. In the IT industry, this allows engineers not only to comply with strict thermal limits for high-performance chips but also to reproduce nanotechnological processes in manufacturing, where minute changes in temperature and atomic movement are crucial for the functionality of the final ICS.

The application of these principles is most evident in nanotechnologies and the manufacturing of SC components. Planar technology remains the primary method for the mass production of SC elements, including processors and microcontrollers. Its continuous improvement is driven by Moore's Law, which necessitates the relentless shrinking of component sizes. Key stages such as lithography, oxidation, and dopant diffusion are governed by thermodynamic variables. For instance, oxidation controls the growth of thin silicon dioxide layers at high temperatures, serving as either a mask or a dielectric gate in MOSFET transistors. The creation of p-n junctions depends on the thermal diffusion of alloying elements within the silicon crystal. By utilising kinetic theory and diffusion models, CP enables the simulation of how dopant atoms distribute within the crystal lattice under high temperatures, allowing for precise control over the depth and concentration profiles that determined electrical characteristics (Martin-Bragado *et al.*, 2006).

Furthermore, molecular physics and thermodynamics are central to the development of cooling systems and thermal design. CP employs energy laws, such as the Joule-Lenz law and the laws of thermodynamics, alongside numerical methods like FEA to create models that optimise processor architecture for reduced heat generation. In high-performance environments, including data centres and mobile devices, numerical simulations of thermal conductivity, convection, and radiation are used to model temperature field distributions. This allows for the optimisation of cooling elements such as thermal pastes, radiators, and heat pipes, the latter of which relied on the physical phenomena of evaporation and condensation to ensure efficient heat dissipation (Tari & Yalcin, 2010).

The search for new materials also relies heavily on materials science and numerical testing of innovative substances. Researchers sought materials with exceptional thermal conductivity, such as graphene or diamond films, to enhance heat dissipation in high-performance computing (Chen *et al.*, 2022). Active research is also focused on SC composites like Gallium Nitride (GaN) and Silicon Carbide (SiC) for power electronics, which maintain superior electrical characteristics at elevated temperatures. Additionally, molecular physics principles are applied to optimise thermal interface materials and develop next-generation memory technologies, such as Resistive Random-access memory (RRAM) and Ferroelectric RAM (FRAM), where the switching of atomic or molecular states is fundamental to data storage. Finally, the principles of molecular physics and thermodynamics underpin the operation of various sensors within SMART systems and the Internet of Things (IoT). Temperature measurement is frequently based on changes in electrical resistance or thermoelectric effects resulting from variations in the intensity of atomic thermal motion. Similarly, gas sensors utilise chemical reactions that generate heat or changes in conductivity upon the adsorption of gas molecules (Goel *et al.*, 2023). Pressure sensors may also employ the thermodynamic properties of gases or the effects of mechanical deformation on a material's electrical properties, further demonstrating the integration of molecular physics into modern monitoring systems.

Electricity and magnetism serve as the fundamental language of ICS, as all processes of information transfer, storage, and processing occur through electric fields, currents, and electromagnetic waves. CP utilises Maxwell's equations and derivative laws, such as those formulated by Ohm, Coulomb, and Faraday, for the numerical modelling of signal propagation, antenna design, data integrity analysis, and interference prevention. Precise modelling and prediction of electromagnetic phenomena are essential for ensuring high speeds, reliability, and the continued miniaturisation of modern electronics. Figure 1 presents the hierarchy of abstraction layers in computing systems, demonstrating the sequential transformation of fundamental physical processes in SCs into logical operations and high-level software.

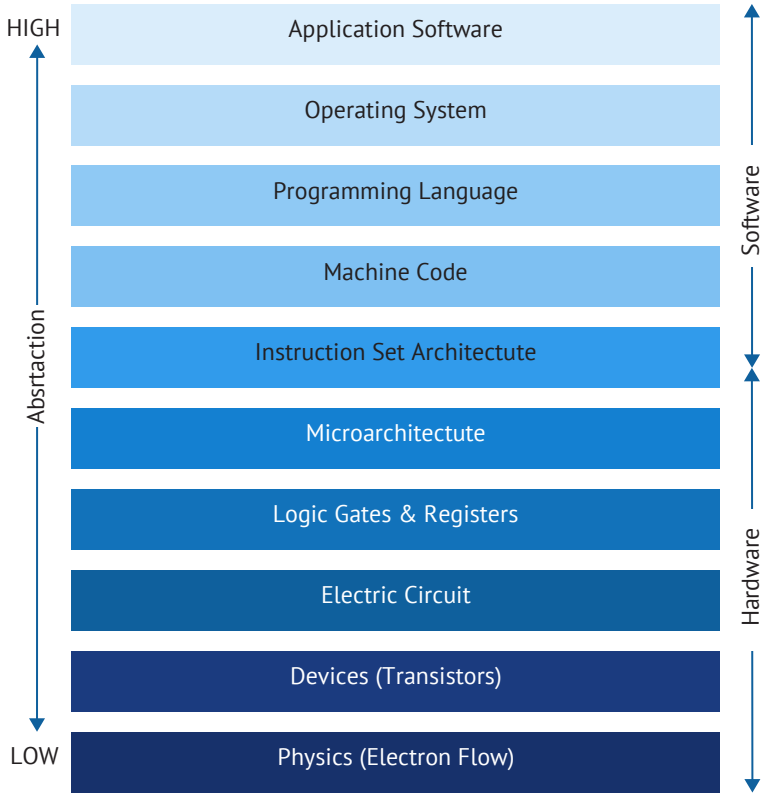


Figure 1. Layers of abstraction in computer systems

Source: compiled by the author

A critical area of application is the maintenance of signal and power integrity. In modern high-speed computers, where switching speeds reach the gigahertz range, electrical signals behave as electromagnetic waves on transmission lines. Consequently, modelling these lines was vital to prevent internal phenomena such as signal reflection and dispersion, which could degrade data quality (Jin *et al.*, 2004). This ensures that electrical pulses arrive at the receiver on time and without distortion. Similarly, power integrity is modelled to ensure the correct operation of inductive elements, such as coils and transformers, and to facilitate noise

filtration through capacitors, maintaining the voltage stability required for reliable chip operation.

Furthermore, electricity and magnetism provide the indispensable foundation for data storage across various media (Ma, 2025). In electrical storage systems, such as Dynamic RAM (DRAM), information is stored as an electric charge in microscopic capacitors, with electrostatics being used to model rapid read/write cycles. Static RAM (SRAM) utilises transistor flip-flops, where electrodynamics describes the ultra-fast switching of currents. For non-volatile storage, such as Flash memory (SSD, EEPROM, EPROM), information is maintained as a charge trapped in a transistor's floating gate, where programming occurs through controlled electric fields that enable electron tunnelling. Emerging technologies, including FRAM and RRAM, use electric fields to change material polarisation or voltage to form conductive filaments, respectively. Magnetic storage remains equally significant; in HDDs, data is encoded via the magnetisation direction of microscopic areas on platters, a process optimised through CP modelling of magnetic fields and magnetoresistive effects. An equally important task in CP is ensuring electromagnetic compatibility and shielding, which allows for the management of undesirable radiation and the susceptibility of devices to external interference. As every electronic device is a potential source and receiver of electromagnetic waves, numerical methods were employed to design effective shielding (Morales *et al.*, 2023). This shielding was based on the mechanisms of reflection by highly conductive metals and absorption by materials with specific dielectric or magnetic losses. Such modelling guaranteed the uninterrupted operation of ICS, from mobile gadgets to complex cyber-physical systems, even in environments with high electromagnetic noise.

Wireless communication technologies, including Wi-Fi, Bluetooth, and 5G/6G, rely entirely on the transmission and reception of electromagnetic waves, with electrodynamics providing the only theory capable of describing antenna operation. CP was used to model field distribution around miniature antennas, optimising radiation patterns and impedance matching based on the electrical and magnetic properties of the structure (Basherlou *et al.*, 2025). Moreover, the introduction of new frequency spectra by engineers ensures the innovativeness and reliability of global communication technologies. Finally, reliable and autonomous power sources are critical for mobile and high-performance systems (Rasolomampionona & Kłos, 2023). CP applies the laws of electrodynamics and thermodynamics to model the electrochemical processes within Li-Ion batteries, enabling the optimisation of capacity, charging speed, and service life. Additionally, the operation of solar cells, used in autonomous IoT sensors and robotics, is based on the photoelectric effect. Here, electrodynamics and optics are utilised to model photon absorption and current generation in SC layers, allowing engineers to maximise energy conversion efficiency and develop flexible power sources.

Optics is a fundamental field of physics that studies the nature of light and its interaction with matter, providing the technological foundation for high-speed data transmission, high-precision manufacturing, and visual information gathering. CP utilises the laws of geometric and wave optics, including reflection, refraction, diffraction, and interference, for the numerical modelling and implementation of complex optical-physical systems. This enables engineers not only to overcome the geometric limitations of miniaturisation in microelectronics but also to create efficient high-performance networks and human-machine interfaces. A primary application of these principles is optical lithography, which is a fundamental process in the manufacturing of microprocessors and memory. This technology operates on the laws of wave optics, using diffraction and interference for the ultra-precise transfer of complex circuit patterns from a photomask onto a silicon wafer. As transistor sizes shrink towards the nanometre scale, as dictated by Moore's Law, the diffraction limit becomes a critical physical boundary. To overcome this, computer engineering employed complex wave optics modelling in the development of innovative systems such as immersion lithography and EUV, which ensured the resolution necessary for further component miniaturisation (Bakshi *et al.*, 2017).

Optoelectronic data transmission represents another vital area, encompassing the conversion of electrical signals into light signals to form the basis of high-speed communication. Fibre-optic networks, which underpin the modern internet and data centres, rely on the phenomenon of total internal reflection at the boundary between the fibre core and cladding. Optics is used to model dispersion and signal loss in long-distance lines, maximising transmission speeds. Beyond telecommunications, fibre optics is utilised in high-speed peripheral connections such as HDMI 2.1 and USB 4, as well as in avionics, space technology, and industrial SMART-networks. Furthermore, optical fibres serve as distributed sensors for monitoring large-scale infrastructure; by analysing how light propagation is affected by external influences, these systems can measure temperature, vibration, and strain in bridges, pipelines, and deep boreholes. The effective operation of these networks also requires the modelling of optical signal generators, such as lasers and LEDs, alongside photodetectors for signal reception (Kim, 2017).

The development of sensors and cameras also relies heavily on optical principles. Geometric optics is used to design lenses and microlens arrays, which are the foundation of computer vision systems in smartphones, UAVs, and autonomous transport. Additionally, LiDAR technologies act as optical distance sensors, relying on the laws of light propagation to measure the time-of-flight of laser pulses. These systems are crucial for 3D scanning and the navigation of autonomous mobile robotics (Zhou, 2022). Finally, optics is essential for

the design of displays and immersive interfaces. In LED and LCD screens, optical modelling optimises brightness, colour accuracy, and viewing angles by analysing the interaction of light with liquid crystals or organic materials. For augmented and virtual reality (AR/VR), optical systems consisting of specialised lenses and prisms are indispensable for the correct display of images on screens positioned close to the eyes. Geometric optics allows for the accurate modelling of distortions and vision correction, ensuring a realistic and comfortable user interface (Song & Wu, 2024).

QM serves as a primary foundation for the modern IT industry, providing the theoretical framework to explain the behaviour of matter at atomic and subatomic levels. Without accounting for quantum effects, such as discrete energy levels and wave-particle duality, the functioning of SC materials – from which all transistors, processors, and memory are constructed – would be impossible. CP utilises quantum principles to model the band structure of materials and the behaviour of electrons in nanoscale structures, effectively defining the physical limits and capabilities of modern computing. While advanced areas such as quantum computing and cryptography remain promising, their full realisation represents a significant technological challenge for contemporary engineering. The understanding of SC and transistors is deeply rooted in QM, which explains how conductivity depends on band structures and how doping creates charge carriers. Modelling based on the Schrödinger equation allowed engineers to predict electron flow in MOSFET transistors, which are the basic building blocks of microchips (Cheng *et al.*, 1999). As transistor sizes shrink to the nanometre level, undesirable quantum effects, such as quantum tunnelling, emerge. This phenomenon describes the probability of an electron overcoming thin dielectric barriers even when its energy is lower than the barrier height. CP employed quantum models to calculate tunnelling probabilities, which is crucial for determining the physical limits of miniaturisation and preventing current leakages that lead to power losses and calculation errors (Khodayari *et al.*, 2026).

Quantum computing represents the most ambitious application of QM, utilising qubits that operate through superposition and quantum entanglement. Unlike classical bits, these quantum phenomena allow for an exponential increase in computational power, enabling the processing of vast numbers of states in parallel (Oukaira, 2025). This paves the way for solving complex problems, including molecular modelling for new materials and the rapid factorisation of large numbers using Shor's algorithm. Closely related to this is the phenomenon of superconductivity, where zero electrical resistance occurs due to the formation of Cooper pairs of electrons. Superconducting circuits containing Josephson junctions are currently a dominant physical principle for creating high-coherence qubits (Huang *et al.*, 2020). Beyond quantum computers, superconductivity offers a path toward energy-efficient classical electronics. Technologies such as Rapid Single Flux Quantum (RSFQ) could enable the creation of ultra-fast processors with minimal thermal losses due to the absence of electrical resistance (Bairamkulov & De Micheli, 2024). Furthermore, quantum metrology utilises superconducting sensors, such as SQUID magnetometers, for ultra-sensitive magnetic field measurements in medical diagnostics and high-precision electronic component verification.

QM also provides the basis for absolutely secure communication through quantum cryptography, specifically Quantum Key Distribution (QKD). Unlike classical methods that rely on mathematical complexity, QKD relies on the fundamental laws of physics. Because any attempt to measure a photon's state inevitably alters it, eavesdropping is immediately detected, making key interception physically impossible (Lütkenhaus, 2019). Finally, Ab Initio modelling methods, particularly DFT, allow for the numerical prediction of physicochemical properties at the atomic level. This fundamental approach accelerates innovation across the IT industry by enabling the discovery of new SC and thermal interface materials with enhanced properties (Kleger & Meunier, 2023). Atomic and nuclear physics study the structure of the atom and its nucleus, as well as processes related to radioactive decay, nuclear reactions, and the interaction of radiation with matter. Although their role in the IT sphere may not be immediately obvious, these fields are essential for ensuring the longevity, reliability, and security of electronic systems, as well as for powering high-performance or autonomous computing complexes.

A primary concern addressed by nuclear physics is the radiation hardness of electronics. High-energy cosmic rays, such as neutrons and muons, interact with the silicon structure of microchips, causing local bursts of ionising radiation. These interactions can lead to Single Event Upsets (SEU), where a single bit state is erroneously switched ($0 \rightarrow 1$ or vice versa). Atomic and nuclear physics are utilised to model these interactions, enabling engineers to develop radiation-hardened microcircuits. This is critically important for aerospace technology, including satellites and Mars rovers, as well as for medical radiation therapy equipment and data centres located at high altitudes (de Aguiar *et al.*, 2025). Furthermore, nuclear physics plays a significant role in the energy sector as it pertains to IT infrastructure. Given that large-scale data centres consume gigawatts of electricity, there is a growing trend toward integrating small modular reactors (SMRs) or utilising nuclear power plants to ensure a stable and sustainable power supply (Vanatta *et al.*, 2024). For autonomous electronics in deep space or long-term missions where solar energy is unavailable, radioisotope thermoelectric generators (RTGs) are employed. The operating principle of RTGs is based on radioactive decay, providing a reliable source of electricity for onboard computers in environments where other power sources would fail.

Additionally, nuclear magnetic resonance (NMR) is applied within the IT industry for more than just medical imaging. While frequently associated with MRI, NMR is used for the non-destructive analysis of the chemical composition and structure of SC materials. This application was vital for quality control in microchip manufacturing, which allowed for high-precision verification of component integrity at the atomic level (Kehayias *et al.*, 2017). Interdisciplinary interconnection**.** The modern IT industry functions not thanks to individual branches of physics, but due to their complex interdisciplinary connection. No complex component (from processor to screen) can be designed using only one theory. Physics of the micro-world (QM, atomic physics) sets the fundamental limits for materials, while classical physics (classical mechanics, electro- and thermodynamics, optics) provides the engineering solutions for their realisation and stable operation in the macro-world.

One of the clearest examples is optoelectronics and displays, where the laws of QM, optics, and molecular physics operate simultaneously. The operation of quantum dots, used in modern QLED screens, is entirely based on QM (the quantum confinement effect, which determines the colour). However, for the correct display of the image, engineers apply optics (reflection and transmission of light) and thermodynamics (managing the thermal regime of the display, which affects longevity and colour). A similar situation is observed with liquid crystals, where electric fields (electrodynamics) control the orientation of molecules (molecular physics), and the result of their work – the reflection of light – is described exclusively by optics. Another example is the reliability and energy efficiency of microcircuits. Here, QM describes the electrical behavior of the transistor, allowing electrons to overcome tunneling barriers. However, the heat generated during the passage of such electrons is described by the laws of thermodynamics. Such a phenomenon must be taken into account by computer engineers to prevent the destruction of the material, whose properties are determined by molecular physics and quantum chemistry (Ab Initio modeling). Moreover, complex cooling systems (heat pipes), whose operation is a direct application of molecular physics phenomena (evaporation and condensation), are used to dissipate this heat.

Robotics and autonomous systems demonstrate a direct synthesis of classical physics. Mechanics is fundamental for calculating the kinematics and dynamics of the robot's motion (forces, inertia, trajectory). But for the robot to "see" the world and make decisions, optics (cameras, LiDAR) and electrodynamics (wireless communication, electric motor control) are needed. For example, an autonomous UAV uses a GPS system (electrodynamics + relativistic mechanics), inertial MEMS sensors (classical mechanics + electrostatics), and its stability in the air is controlled, in particular, by modeling the influence of wind (classical mechanics). The following Figure 2 illustrates a typical cyber-physical system infrastructure, demonstrating the close interconnection between physical processes and the software layer, where real-time sensor data informs the control logic for hardware components.

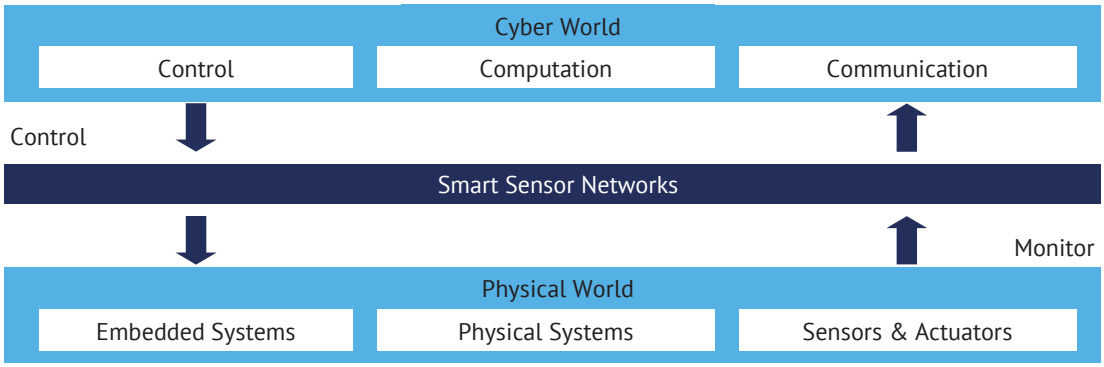


Figure 2. Cyber-physical system infrastructure

Source: compiled by the author

Thus, all IT fields, without exception, exist exclusively at the intersection of these fundamental scientific disciplines, making CP the foundation of all technological innovations. The comprehensive analysis of physical laws within the architecture of ICS reveals a profound shift from traditional hardware engineering towards an integrated computational-physical approach. The results obtained in this study underscore that the evolution of the IT industry is no longer limited by software complexity but by the fundamental physical boundaries of materials, energy, and subatomic interactions. This perspective suggests that CP has moved from being a supplementary tool to becoming the central governing framework for the future of IT. The synthesis of mechanical principles, as explored through the dynamics of high-speed components and MEMS sensors, demonstrates that classical physics remains indispensable even in a digital-first era. The work of M.A. Gharaibeh & B.M.Y. Gharaibeh (2023) regarding vibrational resistance and G.S. Chen *et al.* (2025) concerning HDD read/write head dynamics highlighted a critical point of discussion: as components shrink and

speeds increase, mechanical precision becomes a bottleneck for data integrity. The necessity of modelling these systems using the FEM proved that the mechanical reliability of a system is as vital as its algorithmic efficiency. This alignment between macro-scale structural stability and micro-scale operational precision is a recurring theme throughout the study.

In the field of molecular physics and thermodynamics, the results indicate a fundamental shift from empirical manufacturing to predictive modelling. The analysis of planar technology and dopant diffusion processes, supported by the findings of I. Martin-Bragado *et al.* (2006), demonstrated that adhering to Moore's Law is primarily a thermodynamic challenge. The ability to numerically simulate atomic motion and temperature field distribution ensures a level of precision unattainable by physical experimentation alone. Furthermore, the implementation of advanced cooling solutions, as explored by I. Tari & F.S. Yalcin (2010) and X.-K. Chen *et al.* (2022), confirmed that heat dissipation efficiency and the stabilisation of thermal regimes are currently the principal limiting factors for the advancement of high-performance computing systems.

The transition from thermal constraints to electromagnetic ones is seamless, as seen in the modelling of signal and power integrity. The work of H. Jin *et al.* (2004) remains a cornerstone in understanding how electrical signals behave as electromagnetic waves at gigahertz frequencies. This study's analysis of signal reflection and dispersion confirmed that electricity and magnetism provide the fundamental language of ICS. The discussion of data storage, from DRAM to magnetic HDDs as analysed by Z. Ma (2025), further reinforces the idea that information is, at its core, a physical state – whether it be a trapped charge or a magnetic orientation. The importance of electromagnetic compatibility and shielding, supported by M.A. Morales *et al.* (2023), demonstrated that the isolation of electronic components from external interference is a prerequisite for the reliability of all modern mobile and cyber-physical systems. Furthermore, the results regarding optics and wireless communication highlight the dual nature of light as both a manufacturing tool and a data carrier. The critical boundary represented by the diffraction limit in optical lithography, discussed by V. Bakshi *et al.* (2017), serves as a bridge between classical wave optics and the quantum-scale manufacturing of the future. The application of geometric optics in computer vision and LiDAR technologies, as explored by J. Zhou (2022) and J.-K. Song & S.-T. Wu (2024), demonstrated that the perception of the physical world by autonomous systems is entirely dependent on the accurate modelling of light propagation. This confirms that optics is the primary interface between the digital processor and the physical environment.

The most significant technological shift is observed in the application of QM. The findings regarding quantum tunnelling and the physical limits of miniaturisation, supported by B. Cheng *et al.* (1999) and Z. Khodayari *et al.* (2026), prove that the field has reached a stage where classical logic is insufficient to describe hardware behavior. The analysis of quantum computing and superconductivity (Bairamkulov & De Micheli, 2024; Huang *et al.*, 2020; Oukaira, 2025) defineed the strategic directions for the long-term development of ICS. Furthermore, the reliability of global security based on quantum cryptography, as described by N. Lütkenhaus (2019), demonstrated that the fundamental laws of physics are now the defining basis for ensuring information security during the transition to post-quantum standards. Finally, the role of atomic and nuclear physics in ensuring radiation hardness and providing autonomous power through RTGs (Vanatta *et al.*, 2024; de Aguiar *et al.*, 2025) highlights the extreme environments in which modern ICS must operate. The use of NMR for quality control (Kehayias *et al.*, 2017) and the relativistic corrections required for GNSS systems (Ayso & Kahveci, 2025) serve as reminders that the IT industry must account for the structure of the atom and the curvature of spacetime to maintain modern standards of accuracy.

In conclusion, this article provides a unique synthesis of physical laws across the entire spectrum – from classical and relativistic mechanics to thermodynamics, electromagnetism, and quantum physics – demonstrating their inextricable link to the architecture of modern ICS. Unlike specialised studies that focus on isolated hardware components, this work offers a holistic interdisciplinary framework that identifies physics as the primary driver of IT innovation. The originality of this research lies in its systematic approach to bridging the gap between fundamental scientific theories and practical engineering challenges, such as the thermal wall and quantum tunnelling. By unifying these diverse branches of physics into a single conceptual model, this study establishes a new perspective on the evolution of IT, proving that the future of digital systems is fundamentally governed by the transition to a physics-aware computational paradigm.

CONCLUSIONS

The conducted research, aimed at analysing the interconnection of fundamental physical laws with the hardware and software architecture of ICS and defining directions for their future development, allowed for the formation of a comprehensive vision of the physical foundations of the IT industry. It was established that physical laws are an integral part and the basis for the functioning of ICS. Classical mechanics and electrodynamics underpin robotics, inertial MEMS systems, and LiDAR sensors necessary for autonomous vehicles and aerospace applications. The reliability, longevity, and resistance of such systems to vibrations are determined by mechanical fatigue models. Thermodynamics and molecular physics are key to the effective management of the thermal regime and solving heat

dissipation problems in modern technology, emphasising the importance of CFD analysis and liquid immersion cooling. Thermal conductivity of nanostructures (e.g., PCF-graphene) is becoming critical for passive heat dissipation. Optics and electrodynamics ensure the functioning of high-speed data transmission channels and are also the basis for manufacturing technologies. Specifically, EUV remains key to forming nanometer-level microchips. QM, particularly the phenomenon of tunneling, directly defines the physical limits of further transistor miniaturisation, requiring a transition to new materials (e.g., 2D Janus LiMSSe) and architectures to circumvent these limitations. Atomic and nuclear physics play an important role in ensuring the radiation hardness of electronics, which is critically important for space-based, highly autonomous systems, and are also being considered as a potential power source for industrial and autonomous SMART complexes.

Further research should be focused on critical areas that will determine the evolution of ICS, including the development of reliable and scalable superconducting quantum qubits and the investigation of quantum protocols (e.g., QKD) for secure data transmission. Computational modeling (based on DFT and molecular dynamics methods) is also key for designing new SC materials that will allow circumventing the quantum limits of silicon structure miniaturisation. An important prospect is the integration of physically accurate real-time simulation models (based on N-body dynamics and hydrodynamic models) for extended reality (XR), which will ensure a new level of interaction realism.

ACKNOWLEDGEMENTS

I express my gratitude to Petruk V.P. (Mizoch Lyceum) – my physics teacher, who laid the foundation for a deep understanding of the subject.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Ayso, E., & Kahveci, M. (2025). Impacts of relativistic effects on GNSS signal path and precise point positioning. *Measurement Science and Technology*, 36(6), article number 066318. doi: 10.1088/1361-6501/adddd3.

[2] Bairamkulov, R., & De Micheli, G. (2024). Superconductive electronics: A 25-year review. *IEEE Circuits and Systems Magazine*, 24(2), 16-33. doi: 10.1109/MCAS.2024.3376492.

[3] Bakshi, V., Mizoguchi, H., Liang, T., Grenville, A., & Benschop, J.P. (2017). Special section guest editorial: EUV lithography for the 3-nm node and beyond. *Journal of Micro/Nanolithography, MEMS, and MOEMS*, 16(4), article number 041001. doi: 10.1117/1.JMM.16.4.041001.

[4] Basherlou, H.J., Parchin, N.O., & See, C.H. (2025). Antenna design and optimization for 5G, 6G, and IoT. *Sensors*, 25(5), article number 1494. doi: 10.3390/s25051494.

[5] Cai, Y., *et al.* (2020). Band structure, effective mass, and carrier mobility of few-layer h-AlN under layer and strain engineering. *APL Materials*, 8(2), article number 021107. doi: 10.1063/1.5139664.

[6] Camiola, V.D., Romano, V., & Vitanza, G. (2025). Quantum MEP hydrodynamical model for charge transport. *Journal of Statistical Physics*, 192, article number 20. doi: 10.1007/s10955-025-03395-z.

[7] Chen, G.S., Xu, J., & Hua, W. (2025). Nanoscale dynamics of air-bearing slider in computer hard disk drives. In *Microsystem dynamics* (pp. 99-176). Hoboken: John Wiley & Sons Ltd. doi: 10.1002/9781118848890.ch04.

[8] Chen, X.-K., Hu, X.-Y., Jia, P.-Z., & Xie, G.-F. (2022). First-principles determination of high thermal conductivity of PCF-graphene: A comparison with graphene. *Applied Physics Letters*, 121, article number 182205. doi: 10.1063/5.0123629.

[9] Cheng, B., Cao, M., Rao, R., Inani, A., Vande Voorde, P., & Greene, W.M. (1999). The impact of high-/spl kappa/ gate dielectrics and metal gate electrodes on sub-100 nm MOSFETs. *IEEE Transactions on Electron Devices*, 46(7), 1537-1544. doi: 10.1109/16.772508.

[10] de Aguiar, Y.Q., Wrobel, F., Autran, J.-L., & Alía, G.R. (2025). Introduction to single-event effects. In *Single-event effects, from space to accelerator environments*. Cham: Springer. doi: 10.1088/1361-6501/adddd3.

[11] Ezratty, O. (2023). Perspective on superconducting qubit quantum computing. *The European Physical Journal A*, 59(5), article number 94. doi: 10.1140/epja/s10050-023-01006-7.

[12] Gharaibeh, M.A., & Gharaibeh, B.M.Y. (2023). Analytical evaluation of solder stress in electronic packages subjected to random vibrations. *Mathematical Modelling of Engineering Problems*, 10(4), 1265-1270. doi: 10.18280/mmep.100419.

[13] Goel, N., Kunal, K., Kushwaha, A., & Kumar, M. (2023). Metal oxide semiconductors for gas sensing. *Engineering Reports*, 5(6), article number e12604. doi: 10.1002/eng2.12604.

[14] Hidalgo, M.A. (2022). Quantum hall effects in two-dimensional electron systems: A global approach. *The European Physical Journal Plus*, 137(1), article number 58. doi: 10.1140/epjp/s13360-021-02173-6.

[15] Huang, H.-L., Wu, D., Fan, F., & Zhu, X. (2020). Superconducting quantum computing: A Review. *Science China Information Sciences*, 63, article number 180501. doi: 10.1007/s11432-020-2881-9.

[16] Jin, H., Li, E., Yuan, W., & Li, L. (2004). Parallel FDTD computing for EMC simulation in hihg-speed electronics. *International Journal of Computational Engineering Science*, 5(3), 575-588. doi: 10.1142/S1465876304002575.
[17] Kadri, M.B., Khatri, S.A., & Yousuf, S. (2025). Trajectory tracking control of a planar robotic arm using inverse dynamics and fuzzy gain scheduling: Simulation and experimental validation. *IEEE Access*, 13, 186736-186759. doi: 10.1109/ACCESS.2025.3626418.
[18] Kalavathi, D. (2025). Modelling and optimisation of electromagnetic interference (EMI) in complex electronic system. *International Journal on Science and Technology*, 16(4), 1-10. doi: 10.71097/IJSAT.v16.i4.9770.
[19] Kehayias, P., *et al.* (2017). Solution nuclear magnetic resonance spectroscopy on a nanostructured diamond chip. *Nature Communication*, 8, article number 188. doi: 10.1038/s41467-017-00266-4.
[20] Kempe, V. (2011). *Inertial MEMS: Principles and practice*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511933899.
[21] Khodayari, Z., Karimi R., & Ghobadi, N. (2026). Computational design of sub-5 nm high-performance double-gate FETs based on two-dimensional Janus LiMSSe (M = Al, Ga, In) monolayers. *Applied Surface Science*, 716, article number 164730. doi: 10.1016/j.apsusc.2025.164730.
[22] Kim, H.-S., Kim, Y.-G., Jang, D.-M., Jang, J.-W., Lee, S.-Y., & Kim, H.-S. (2025). Prediction of fatigue life of semiconductor package under thermal cycling: Combined effect of package design and solder materials. *Materials & Design*, 254, article number 114059. doi: 10.1016/j.matdes.2025.114059.
[23] Kim, T.-H. (2017). Analysis of optical communications, fiber optics, sensors and laser applications. *Journal of Machine and Computing*, 3(2), 115-125. doi: 10.53759/7669/jmc202303012.
[24] Kleger, A., & Meunier, V. (2023). Density functional theory study of the structural, electronic, mechanical, and thermal properties of $Hf_6Ta_2O_{17}$. *Materials Today Communications*, 34, article number 105065. doi: 10.1016/j.mtcomm.2022.105065.
[25] Kumari, P. (2024). Quantum tunneling effects in ultra-scaled MOSFETs: A theoretical perspective on device miniaturization limits. *International Journal of Physics and Applications*, 6(1(B)), 139-143. doi: 10.33545/26647575.2024.v6.i1b.153.
[26] Lütkenhaus, N. (2019). Theory of quantum key distribution (QKD). In D. Bruß & G. Leuchs (Eds.), *Quantum information: From foundations to quantum technology applications* (pp. 353-367). Hoboken: John Wiley & Sons Ltd. doi: 10.1002/9783527805785.ch16.
[27] Ma, Z. (2025). Traditional and emerging memory technologies: Principles, applications and future trends. In *Proceedings of the 2025 international conference on electronic and electrical information communications (EEIC 2025)* (pp. 439-452). Amsterdam: Atlantis Press. doi: 10.2991/978-94-6463-864-6_41.
[28] Martin-Bragado, I., Tian, S.K., Johnson, M., Castrillo, P., Pinacho, R., Rubio, J., & Jaraiz M. (2006). Modeling charged defects, dopant diffusion and activation mechanisms for TCAD simulations using kinetic Monte Carlo. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 253(1-2), 63-67. doi: 10.1016/j.nimb.2006.10.035.
[29] Mohammed, N.M., El-Maghlany, W.M., Elhelw, M., & Abdelaziz, A.H. (2025). Performance improvement of high-density data center via two-phase liquid immersion cooling. *Journal of Thermal Analysis and Calorimetry*, 150, 2907-2924. doi: 10.1007/s10973-025-14006-0.
[30] Morales, M.A., Henry, T.C., & Salamanca-Riba, L.G. (2023). Model of electromagnetic interference shielding effectiveness for a multifunctional composite containing carbon-fiber-reinforced polymer and copper mesh layers. *Carbon*, 212, article number 118179. doi: 10.1016/j.carbon.2023.118179.
[31] Mubashr, A. (2025). Density functional theory: A quantum mechanical framework for novel materials design. *Journal of Engineering, Science and Technological Trends*, 2(3). doi: 10.64060/jestt.v2i3.1.
[32] Naveed, F.N.U., & Dix, J. (2025). A radiation-hardened 4-bit flash ADC with compact fault-tolerant logic for SEU mitigation. *Electronics*, 14(21), article number 4176. doi: 10.3390/electronics14214176.
[33] Oukaira, A. (2025). Quantum hardware devices (QHDs): Opportunities and challenges. *IEEE Access*, 13, 98229-98241. doi: 10.1109/ACCESS.2025.3576216.
[34] Pal, S., & Ray, B.C. (2020). *Molecular dynamics simulation of nanostructured materials: An understanding of mechanical behavior (1st ed.)*. Boca Raton: CRC Press. doi: 10.1201/9780429019845.
[35] Patsis, P.A., & Okalidis, P. (2025). Normal spiral grand-design morphologies in self-consistent n-body models. *Galaxies*, 13(6), article number 132. doi: 10.3390/galaxies13060132.
[36] Pendse, P. (2024). An overview on LiDAR for autonomous vehicles. *Zenodo*. doi: 10.5281/zenodo.10992392.
[37] Rasolomampionona, D.D., & Kłos, M. (2023). Energy storage systems and their role in smart grids. In *Smart grids technology and applications*. London: IntechOpen. doi: 10.5772/intechopen.103945.
[38] Song, J.-K., & Wu, S.-T. (2024). Optical materials and applications for AR and VR display systems. *ACS Applied Optical Materials*, 2(7), 1245-1246. doi: 10.1021/acsaom.4c00226.
[39] Sung, N.-J., Ma, J., Hor, K., Kim, T., Va, H., Choi, Y.-J., & Hong, M. (2025). Real-time physics simulation method for XR application. *Computers*, 14(1), article number 17. doi: 10.3390/computers14010017.

[40] Tari, I., & Yalcin, F.S. (2010). CFD analyses of a notebook computer thermal management system and a proposed passive cooling alternative. *IEEE Transactions on Components and Packaging Technologies*, 33(2), 443-452. doi: 10.1109/TCAPT.2010.2044505.
[41] Vanatta, M., Patel, D., Allen, T., Cooper, D., & Craig, M.T. (2024). Technoeconomic analysis of small modular reactors decarbonizing industrial process heat. *Joule*, 8(2), 542-552. doi: 10.1016/j.joule.2024.02.001.
[42] Wang, S., Cao, Y., Zheng, X., & Zhang, T. (2022). Collision-free trajectory planning for a 6-DoF free-floating space robot via hierarchical decoupling optimization. *IEEE Robotics and Automation Letters*, 7(2), 4953-4960. doi: 10.1109/LRA.2022.3152698.
[43] Yunping, L., Xijie, H., Yonghong, Z., & Yukang, Z. (2018). Dynamic stability and control of a manipulating unmanned aerial vehicle. *International Journal of Aerospace Engineering*, 2018(9), article number 3481328. doi: 10.1155/2018/3481328.
[44] Zhou, J. (2022). A review of LiDAR sensor technologies for perception in automated driving. *Academic Journal of Science and Technology*, 3(3), 255-261. doi: 10.54097/ajst.v3i3.2993.

Фізичні аспекти IT-індустрії: від програмних архітектур до кіберфізичних систем

Тарас Ляшук
Кандидат фізико-математичних наук, старший викладач
Рівненський державний гуманітарний університет
33028, вул. Степана Бандери, 12, м. Рівне, Україна
https://orcid.org/0000-0002-2242-7537

Анотація. Функціонування сучасних обчислювальних систем базується на складному взаємозв'язку апаратного забезпечення та прикладних фізичних явищ. Актуальність дослідження визначається фундаментальною залежністю технологічних інновацій в IT-індустрії від законів фізики, які водночас визначають як межі (наприклад, квантове тунелювання в нанометрових транзисторах), так і нові можливості (наприклад, квантова суперпозиція у кубітах). Метою роботи було проаналізувати взаємозв'язок фундаментальних фізичних законів з апаратно-програмною архітектурою обчислювальних систем та визначати напрямки майбутнього розвитку. Методологія дослідження базувалася на системному підході, що включав контент-аналіз високоякісних наукових даних, структурно-функціональний аналіз фізичних кластерів та теоретичний синтез міждисциплінарних зв'язків, підкріплених логіко-графічним моделюванням. Основні результати дослідження виявили, що квантова механіка і термодинаміка задають фізичні межі для мініатюризації транзисторів та керують тепловим режимом мікросхем, а оптика є основою для високошвидкісної передачі даних і літографії мікрочипів. Класична механіка та електродинаміка забезпечують функціонування кіберфізичних систем, включаючи автоматизовані систем керування, роботизовані комплекси та інфраструктуру бездротового зв'язку. Проаналізовано, що молекулярна фізика та моделювання нанорозмірних структур відіграють вирішальну роль у проєктуванні нових напівпровідникових матеріалів і прогнозуванні довговічності компонентів на основі відповідних комп'ютерно-фізичних моделей. Показано, що ядерна та атомна фізика відіграють ключову роль у забезпеченні радіаційної стійкості електроніки і розглядаються як потенційне джерело живлення для високоавтономних обчислювальних комплексів. Особлива увага приділена перспективним напрямам, зокрема надпровідності, квантовим обчисленням та квантовій криптографії. Практична цінність роботи полягає у формуванні комплексного погляду на фізичні основи обчислювальних систем, що є критично важливим для комп'ютерних та програмних інженерів, науковців при розробці нових матеріалів, проєктуванні високопродуктивних і енергоефективних мікросхем, а також для створення надійних автономних та квантових обчислювальних систем

Ключові слова: комп'ютерна фізика; нанотехнології; комп'ютерне моделювання; планарна технологія; квантові обчислення; енергоефективність; цілісність сигналу

BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 2. 2026

UDC 621.391.82:621.396.96
DOI: 10.62660/bcstu/2.2026.109

Method for determining intentional interference type in conditions of uncertainty of interference environment based on known models

Hryhorii Radzivilov

PhD in Technical Sciences, Associate Professor
Kruty Heroes Military Institute of Telecommunications and Information Technology
01011, 45/1 Knyaziv Ostrozkykh Str., Kyiv, Ukraine
https://orcid.org/0000-0002-6047-1897

Dmytro Pavliuk*

Adjunct
Kruty Heroes Military Institute of Telecommunications and Information Technology
01011, 45/1 Knyaziv Ostrozkykh Str., Kyiv, Ukraine
https://orcid.org/0000-0001-8461-3899

Abstract. The study aimed to theoretically substantiate a method for the identification of intentional interference types in an uncertain interference environment for software-defined radio systems by using known interference models and formalised analysis of the communication channel state vector. The methodological basis included window calculation of the matched feature vector, their normalisation, signature coding, metric and probabilistic comparison with a template library, threshold rejection of uncertain cases, and model-simulation verification in the streaming architecture of software-defined radio. As a result, a formalised feature space and signature profiles of five basic classes of interference were formed: impulse, broadband noise, harmonic, modulation, and low-visibility – as stable multi-parameter patterns of common signal indicator changes that ensure class separation in different signal-to-noise ratio modes, bit error probability variability, and spectral power density morphology, and are reproduced in repeated runs with unchanged parameters. The method is implemented in the form of a computational pipeline with a sequence of stages "state vector → event detection → segmentation → signature description → dimensionality reduction → comparison → decision", where resource-intensive operations are activated only in the presence of an event trigger, and the interfaces between the blocks record a transition from observations to a compact impact signature. The resulting classification rules with membership thresholds and a controlled rejection procedure have been formulated, which transfers new signatures to the accumulation mode and incremental supplementation of the database with subsequent confirmation of profile stability on repeated observations. Simulation testing on a series of parameterised scenarios showed the reproducibility of signature profiles and classification decisions under fixed settings. The practical significance of the study results is determined by the possibility of implementing the method by developers and integrators of embedded software-controlled radio systems and radio monitoring systems for streaming determination of the type of intentional interference based on signature profiles with threshold rejection of unknown cases and controlled replenishment of the signature library

Keywords: channel state vector; time-frequency descriptors; computational pipeline; signature coding; reject option

Article's History: Received: 27.10.2025; Revised: 06.04.2026; Accepted: 18.05.2026; Published: 26.06.2026

Suggested Citation:
Radzivilov, H., & Pavliuk, D. (2026). Method for determining intentional interference type in conditions of uncertainty of interference environment based on known models. *Bulletin of Cherkasy State Technological University*, 31(2), 109-125. doi: 10.62660/bcstu/2.2026.109.

*Corresponding author (pavliuk-d-20@outlook.com)

Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/)

INTRODUCTION

The research relevance is determined by the fact that Software-Defined Radio (SDR) in real conditions of radio-electronic warfare loses its operability not only due to background noise, but primarily due to specific scenarios of deliberate interference, which manifest themselves in various combinations of degradation of the Signal-to-Noise Ratio (SNR), Received Signal Strength Indicator (RSSI), Bit Error Rate (BER) and Power Spectral Density (PSD). Under such conditions, it is not enough to simply record the fact that interference is present – it is critical to algorithmically determine its type based on measured parameters and time-spectral characteristics. Interference classification is considered a computational data processing task with a procedure for rejecting undefined classes and the possibility of expanding signature descriptions, which is especially relevant for SDR platforms with limited computational resources and strict processing delay requirements.

In studies of interference immunity in mobile radio networks, S. Boholii *et al.* (2022) showed that adaptive beamforming provides a significant reduction in interference due to spatial signal selection. The results confirmed the effectiveness of spatial adaptation, but the classification interpretation of the interference type in algorithmic form was not considered. T. Hurskyi *et al.* (2025) provided a quantitative assessment of the effectiveness of radio-electronic suppression of control channels of First-Person View (FPV) unmanned platforms. The channel degradation indicators described the effects of the interference but did not form a procedure for computational recognition of the nature of the interference signal. A description of models and technical methods for generating radio interference was presented by V. Zylevich & S. Svakha (2025), which reveals the energy and spectral properties of various types of interference in telecommunications systems. This presentation provided the basis for parameterised modelling of interference and the formation of feature templates in simulation environments. The algorithmic logic of step-by-step situational control in conflict radio-electronic interaction was considered by V. Kantsedal & A. Mogyla (2025), where the response process is presented as a sequence of computational steps. However, the mechanism for rejecting unknown types of interference and forming new classes in the classification model is not formalised.

The problem of interference recognition in open-set conditions has been studied in detail by H. Han *et al.* (2022), who introduced an open-set formulation for the jamming pattern recognition problem. The paper proves that the threshold decision rule reduces the risk of misclassification of unknown interference patterns and increases the reliability of automatic analysis. A systemic risk approach to intentional electromagnetic interference was proposed by N. Davies *et al.* (2025), where interference is considered as a factor in the interaction of subsystems of complex radio-electronic complexes. This interpretation emphasised the feasibility of embedding interference-type determination algorithms into the computational architecture of radio systems as a separate analytical module. Methods for interference recognition based on multi-domain features were proposed by P. Wang *et al.* (2022), where the combination of several feature spaces increases the accuracy of classification in wireless networks. The results confirmed the effectiveness of the multidimensional representation of the channel state as an input vector for the recognition model.

Automatic classification of interference signals in cognitive radio systems of unmanned platforms was investigated by A. Krayani *et al.* (2022), where time-frequency representations were used as robust descriptors of the interference type. The suitability of compact time-spectral features for practical recognition algorithms was demonstrated. A spectrogram-oriented approach to detecting and classifying interference in Orthogonal Frequency Division Multiplexing (OFDM) systems for unmanned aerial vehicles is presented by Y. Li & M. Pawlak (2022), where machine learning (ML) is combined with specialised feature selection. The results showed that the alignment of the feature structure and the classification model determines the robustness of recognition at low SNR values. Practical methods for detecting and ML-classifying interference in satellite navigation signals were discussed by J. Rijnsdorp *et al.* (2023), where the emphasis is on the suitability of algorithms for implementation in receiving devices. This approach emphasised the importance of resource efficiency, controlled complexity, and scalability of classification solutions.

A synthesis of the above sources revealed a research niche: existing works either demonstrate ways to improve noise immunity or describe interference classification, but do not sufficiently combine these lines into a formalised method for determining the type of intentional interference in conditions of uncertainty of the interference environment as a software-implemented data processing system for SDR. There remains a need for a clearly defined data structure (channel state vector as a multidimensional feature space), a reproducible algorithmic pipeline (collection → pre-processing → feature extraction → classification → reject → signature database update), as well as explicitly stated complexity and resource constraints that ensure practical applicability for SDR platforms. These requirements set the framework for the further development of the proposed method and its simulation validation in a software environment, focused on application in further research and development of a knowledge base of task templates. The study aimed to theoretically justify a method for determining the type of intentional interference in conditions of uncertainty of the interference environment for software-defined radio systems based on the use of known interference

models and formalised analysis of the communication channel state vector.

MATERIALS AND METHODS

The research was conducted as theoretically and algorithmically oriented and aimed at constructing a reproducible computational scheme for determining the type of intentional radio interference in conditions of an uncertain interference environment for SDR. The material basis was formed by formalised parametric descriptions of the radio channel state, presented as time series of measured indicators and their derivatives in the form of feature vectors. The basic state vector included the signal-to-noise ratio (SNR), received signal strength indicator (RSSI), bit error rate (BER), spectral power density (PSD), as well as time statistics (mean values, variances, rates of change, stationarity indicators) and derived time-frequency descriptors formed from short-time transformations and local spectral estimates. All parameters were interpreted as a unified multidimensional feature space of the channel state, in which the type of interference is manifested not by a single indicator, but by a combination of coordinated changes in energy, error, and spectral-structural characteristics, which makes it possible to describe the impact and further classification. To ensure the reproducibility of algorithmic solutions and the possibility of re-testing in different configurations, feature formation was conducted in a fixed computing mode: with a given sampling frequency, sliding window length for SNR/BER/RSSI estimation, PSD estimation parameters (window type, overlap, frequency resolution) and rules for matching temporal and spectral features in a common vector. The rules for normalising features (scaling to comparable ranges and resistance to changes in absolute signal levels) were recorded separately, since it is normalisation that determines the correctness of similarity metrics for comparing signatures in the knowledge base.

In addition, a formalised taxonomy of types of intentional interference and corresponding signature manifestations in the multidimensional feature space of the channel state was used as the material basis for the study. Within the model corpus, impulse, broadband noise, harmonic, modulation, and structured low-probability of intercept/low probability of detection (LPI/LPD) influences were considered, for which coordinated profiles of changes in SNR, RSSI, BER and PSD, as well as time-spectral patterns in the form of local energy maxima, spectral peaks, sidebands and quasi-periodic structures were set for the LPI/LPD class of low-probability-of-intercept/low-probability-of-detection influences. For the reproducible formation of these profiles in the model body, controlled parameterised synthetic implementations of the useful signal and interference effects in the complex I/Q baseband were used. BPSK, QPSK, and 16-QAM with pulse shaping via RRC filtering, as well as OFDM with a cyclic prefix as a control case of a structured spectrum, were used as reference models of the useful signal to verify the robustness of the features to different "normal" spectral morphologies of transmission. Low-power interference (LPI/LPD) was generated in forms that minimize contrast at average SNR/RSSI, leaving reproducible local time-frequency invariants. These include noise-like DSSS signals based on pseudo-random sequences and FHSS influences, where a short time is spent at a frequency, resulting in short local energy maxima in the TF representation instead of stable peaks in the global PSD. Additionally, low-energy chirp/LFM components with a characteristic "sloped" trace in the spectrogram and burst signals with a low duty cycle provide episodic TF anomalies and increase non-stationarity without prolonged degradation of energy averages.

The parameters of the influences were set in a normalised form relative to the sampling frequency and the length of the analysis window (relative bandwidth, frequency sweep rate, dwell time at frequency, duty cycle), and the amplitude ratios were adjusted through SNR scenarios so that for LPI/LPD influences, the main class separation was provided by spectral, entropy, and time-frequency descriptors, rather than absolute power levels. The interference signature was interpreted as a reproducible vector representation of a channel state fragment formed from energy, statistical, and time-frequency descriptors, including estimates of spectral entropy, occupied bandwidth, pulse structure regularity, BER variability, and RSSI change rates. To ensure comparability of signatures between scenarios, coordinated signature coding was used with a reduction in the dimensionality of the feature space using autoencoder or perceptron transformations, which made it possible to obtain compact latent vectors suitable for metric comparison. Thus, the research materials included not only time series of channel parameters and synthetically generated signals/interference, but also formalised signature descriptions of typical interference effects and rules for their vector representation, which were used as a basic information set for further algorithmic classification and incremental expansion of the knowledge base.

The methodological part was based on a combination of system analysis, feature modelling and algorithmic signal classification and was implemented as a formalised computational pipeline: "channel state vector → interference event detection → signature coding → metric comparison → classification decision → rejection or database update". The channel state vector was defined as and was subject to normalisation and time smoothing in sliding windows to reduce the impact of random fluctuations and ensure the stability of subsequent estimates, with the specification of the feature composition being conducted in the form (1):

$$x(t) = [SNR(t), RSSI(t), BER(t), PSD(t), H(t), \sigma^2(t)] \in R^n, \quad (1)$$

where $H(t)$ – channel impulse response; $\sigma^2(t)$ – signal dispersion.

Detection of the interference event was implemented as a multi-criteria threshold procedure incorporating combined changes in SNR, BER, RSSI and spectral indicators, followed by signal segmentation using the adaptive window method with the selection of quasi-stationary fragments. For each fragment, a set of temporal, spectral, and time-frequency features was calculated, including energy distributions, spectral entropy, bandwidth estimates, peak structure parameters, and regularity indicators. Signature coding was implemented by converting the primary feature space into a compact vector representation using dimension reduction and machine learning methods implemented in scikit-learn to reduce the computational load and improve the stability of similarity metrics for SDR platforms with limited resources. Signature comparison was performed using cosine similarity and Euclidean distance, and the decision was made based on the maximum estimate of membership in known classes.

To account for the uncertainty of the interference environment, a rejection mechanism was used, whereby the input vector was classified as an unknown type of interference provided that the maximum estimate of membership in the set of known classes did not exceed the threshold value . In the case of rejection, a new signature record with metadata was formed, and the procedure for incremental replenishment of the knowledge base was initiated after confirmation of the stability of repeated observations. To ensure the scalability and consistency of the signature library in multi-node operating scenarios, the method provided for a two-level architecture of interaction between edge software-defined radio devices and the cloud environment, where local SDR platforms perform initial detection, segmentation, feature formation, and classification decisions in real time with minimal delay, while the cloud level is used as a centralised repository for the global signature database, a service for synchronising updates between devices, and a computing circuit for resource-intensive operations, including retraining dimensionality reduction models or classifiers on expanded datasets, as well as in-depth analysis of ambiguous or rejected cases that require more templates, repeated runs, and more complex validation procedures than are available on edge devices.

Algorithmic suitability was assessed by analysing computational complexity and resource requirements, in particular the complexity of matching a signature with a database of order , where – dimension of the feature vector and – number of templates in the database. MATLAB (Version R2024a), Python (Version 3.12) with NumPy, SciPy and scikit-learn libraries, as well as GNU Radio (Version 3.10.0.0) for SDR-oriented simulation scenarios were used as the modelling and software environment. MATLAB was used for controlled numerical experiments with signal/noise generation and verification of the correctness of feature calculations and threshold rules; Python was used for reproducible data processing implementations, batch data set formation, and parametric runs; NumPy and SciPy were used as tools for numerical calculations, transformations, and spectral estimates; scikit-learn was used for the implementation of classification procedures, threshold calibration, and verification of decision consistency; GNU Radio was used for modelling the SDR streaming architecture, where collection, pre-processing, feature formation, and classifier decisions are interpreted as a sequence of software blocks with delay and computational complexity constraints. Calibration of membership thresholds τ_sim, τ_dist and τ_prob was performed as a separate optimisation procedure on a validation subset of scenarios with known interference classes and a control set of "unknown" profiles.

The optimisation goal was to simultaneously maximise the proportion of correct classification of known classes (Accuracy_known) and minimise the proportion of false classification of unknown influences (False Acceptance Rate_unknown), which was formalised through a multi-criteria function of the form $J(\tau) = w_1 \cdot F1\_known - w_2 \cdot FAR\_unknown$, where F1_known is the F1 measure for known classes, FAR_unknown is the proportion of unknown signatures mistakenly assigned to existing templates, and w_1 and w_2 are the weighting coefficients for stable typing and protection against forced classification. For the cosine similarity metric scheme, the threshold τ_sim was determined by the maximum $J(\tau)$ in the search process within the permissible interval [0,1], for the Euclidean distance τ_dist – as the minimum distance value that ensures the specified True Positive Rate_known level with the restriction FAR_unknown $\leqslant \varepsilon$, and for the probabilistic classifier τ_prob – based on the analysis of ROC curves and the selection of the point that maximises the Jude index (TPR – FPR). Additionally, the stability of the optimal τ was verified for variations in SNR and window analysis parameters; the threshold was considered acceptable only if the decision was invariant within the acceptable range of scenarios. This procedure ensured a formally justified selection of thresholds that balances sensitivity to known classes and resistance to unknown signatures in conditions of uncertain interference. Within the framework of model-simulation validation, the delay, operational memory, and scalability of the signature database in streaming mode were quantitatively assessed, as these parameters determine the practical suitability of the conveyor for implementation on edge-SDR devices. The evaluation was performed for two reference computing profiles: "desktop-contour" for control reproducibility (x86-64 class CPU) and "edge-contour" for limited resources (ARM class), in a single streaming analysis configuration with a sliding window and a conditionally activated extended processing branch.

RESULTS

Formalisation of multidimensional feature space and normalised radio channel state vector

Formalisation of the feature space and channel state vector showed that heterogeneous radio channel

observation parameters can be reduced to a consistent multidimensional representation suitable for signature identification of intentional interference. As a result of computational processing, the energy, error, spectral, temporal, and time-frequency characteristics were reduced to a single normalised state vector x(t), formed in fixed sliding windows and with consistent scaling rules. The formation of x(t) was conducted following representation (1), where the components SNR(t), RSSI(t), BER(t), PSD(t), H(t) and $\sigma^2(t)$ after normalisation were used as the basic input for feature grouping and subsequent signature coding. This reduction eliminated the heterogeneity of the original scales and ensured the metric comparability of signatures between different interference scenarios. The generalised result of feature structuring and their role in signature description is presented in Table 1.

Table 1. Composition and grouping of features in the channel state vector for signature identification of interference

Group of characteristics	Indicators	Computational descriptors	Role in signature description
Energetic	SNR, RSSI	Average in window, local deviations, rate of change	Record signal energy degradation and abnormal level jumps
Error-based	BER	Moving average, variability, threshold exceedances	Reflect the impact of interference on reception integrity
Spectral	PSD	Bandwidth, spectral peaks, uneven distribution	Can distinguish between narrowband, broadband, and structural influences
Temporal	Time series of parameters	Dispersion, autocorrelation, stationarity	Detect impulsivity and non-stationary modes
Time-frequency	Local TF representations	Spectral entropy, local energy maxima	Hidden and subtle structural patterns are identified
Structural	Derived features	Regularity, quasi-periodicity, impulse indices	Forming invariant signatures of interference types

Notes: local TF representations are local time-frequency representations of a signal obtained in short sliding windows as a distribution of energy over time and frequency (e.g., based on short-time Fourier transform or wavelet transform) to highlight short-term and subtle spectral-structural patterns

Source: compiled by the authors using MATLAB R2024a and Python 3.12

Analysis of the generalised grouping in Table 1 showed that the key condition for reproducible signature identification is the functional structuring of heterogeneous channel parameters into coordinated feature blocks, rather than their simple enumeration. Energy features (SNR, RSSI) in this logic serve as a primary indicator of interference events, but do not in themselves provide reliable typification, since different classes of interference can produce similar levels of degradation. The bit error rate (BER) serves as a functional confirmation of the impact on reception and can be used for short-term peak disturbances to be distinguished from stationary changes in the channel. Spectral features (PSD and derivatives) form the basic morphological description of interference, capturing the differences between broadband noise effects, narrowband harmonic components, and more complex modulation structures. Time statistics and time-frequency descriptors enhance discrimination in scenarios with weak "visibility" in SNR/RSSI and enable the detection of structural patterns, in particular for LPI/LPD influences, where local TF features, entropy and regularity invariants become decisive.

The study also established that the correctness of metric signature comparison is determined not only by the composition of features, but also by fixed rules of normalisation and window matching, as they reduce sensitivity to scale and random fluctuations and ensure comparability between scenarios. As a result, formalisation of the feature space as a coordinated signature database has been recorded, in which each feature block has a functional role in the final decision. Energy features (SNR/RSSI and derivatives) provide primary detection of reception deterioration, false features (BER in windows, variability) reflect the impact on transmission quality and separate "noise" deterioration from structured interference, spectral features (PSD, bandwidth, peak structure, entropy) reproduce the morphology of interference in the frequency domain, and temporal and time-frequency descriptors (impulsiveness, regularity, non-stationarity) provide structural confirmation of the class. Normalisation and window matching are critical for correct metric comparison, as they translate signature similarity into dependence on the type of influence rather than absolute levels and random fluctuations. As a result, the formed space supported two key modes of the method: stable classification of known classes and justified rejection of undefined cases when the profile does not match any template.

Signature multiparametric profiles of the main types of intentional radio interference

Generalisation of the results of signature coding showed that for different classes of intentional interference in a multidimensional feature space, a reproducible profile of compatible changes in energy, error, spectral and time-structural indicators is formed. In contrast to isolated channel degradation indicators, it

is the stable combinations of features in sliding windows that ensure class separability and suitability for template matching. As a result of model-simulation runs and signature coding, a generalised set of profiles for five basic types of intentional interference has been formed, presented in Table 2.

Table 2. Generalised signature characteristics of impulse, broadband noise, harmonic, modulation and LPI/LPD interference

Type of interference	Energy indicators (SNR, RSSI)	BER profile	Spectral characteristics (PSD)	Time and TF indicators	Stable signature pattern
Impulse	Short deep SNR dips, sharp RSSI jumps	Peak BER bursts	Broadband energy emissions	Short pulses, high non-stationarity	Rare or serial energy peaks + BER bursts in narrow windows
Broadband noise	Sustained reduction in SNR	Consistently elevated BER	Uniformly increased PSD across the band	Temporal stationarity	Flat spectral profile + prolonged degradation
Harmonious	Moderate decrease in SNR with relatively stable RSSI	Moderate growth in BER	Narrow spectral peaks	Frequency	Dominant narrow line + regularity in time
Modulational	SNR fluctuations	Non-uniform BER	Sidebars, complex PSD structure	Quasi-periodic structures	Spectral component packets + unstable BER
LPI/LPD	Weak SNR/RSSI changes	Low or fluctuating BER	Weak local patterns	Hidden regularity	Low-energy TF patterns + entropy deviations

Source: compiled by the authors using GNU Radio 3.10.0.0 and Python 3.12

Analysis of the generalised profiles in Table 2 showed that the separation of types is ensured not by the extremity of individual parameters, but by the configuration of their joint behaviour. Pulse influences are reliably separated by short-duration energy and false peakiness, while broadband noise influences are separated by a combination of stationary spectral "flatness" and persistently elevated BER. Harmonic and modulation interference are separated primarily by spectral morphology: isolated narrow peaks versus a multi-component structure with sidebands. Insignificant LPI/LPD effects do not form distinct energy anomalies, but demonstrate reproducible time-frequency and entropy deviations, which makes their identification possible only in an extended feature space. The conducted signature typology formed a reproducible correspondence between "interference class → stable multi-parameter pattern", i.e., it transferred the description of interference from the level of general characteristics to the level of specific coordinated changes in the feature space suitable for template comparison.

Within this logic, each class is recorded not by a single marker, but by a configuration of interrelated manifestations in SNR, RSSI, BER, PSD and derived time-spectral descriptors, thereby reducing the risk of class substitution by random fluctuations of a single indicator. The profiles obtained can be interpreted as "signature trajectories" of the channel state: for impulse influences, eventfulness in time dominates with short local disturbances and a characteristic increase in BER instability; for broadband noise – relative stability in time with uniform PSD "filling" and systematic deterioration of error characteristics; for harmonic – narrowband peak structure in PSD and regularity of temporal manifestations with less pronounced changes in average RSSI; for modulation – a combination of sidebands and more complex spectral morphology with quasi-stable temporal fragments and wave-like BER behaviour; for LPI/LPD – low contrast of energy indicators with the simultaneous appearance of weak repetitive patterns in time-frequency indicators, which become key for classifying the impact as a structured, low-visibility class.

A significant result was that typology provides a "minimally sufficient" description of a class in terms of coordinated changes in different groups of characteristics: energy indicators provide the fact and direction of degradation, false indicators provide its practical consequences, spectral indicators provide the form of intervention, and time-frequency indicators provide structural organisation and repeatability. Thanks to this, signature templates become suitable for metric comparison both in the mode of stable classification of known classes and in the mode of incremental expansion of the database: when a new implementation demonstrates a repeatable coordinated profile that does not systematically coincide with existing templates, it can be defined as a candidate for a new signature without the need for complete retraining of the core. Thus, the former typology acts as an operational 'language' for the knowledge base: it provides comparable, reproducible, and scalable descriptions of interference that directly support template comparison, threshold rejection of undefined cases, and controlled replenishment of the signature library as new classes of influence appear.

Computational SDR pipeline for streaming processing and real-time decisions

Within the scope of the results obtained from algorithmic formalisation, the method for determining the type of interference is presented not as a set of separate procedures, but as a complete reproducible computational pipeline in which all processing stages are arranged in a sequential chain with fixed input and output data interfaces. This structuring shifts the signal analysis from the level of "individual operations" to the level of a flowchart decision-making scheme suitable for implementation in SDR-oriented systems under conditions of limited resources and delays. To capture the reproducible logic of data flow through the method rather than a descriptive list of operations, the algorithm was reduced to a structural block diagram with two fundamentally different modes of operation: routine monitoring without activating resource-intensive procedures and extended processing, which is triggered only after detecting a fault event according to agreed threshold criteria.

As a result of the architectural generalisation of the conveyor, the method was found to be suitable for two-level implementation with the distribution of calculations between the edge and cloud levels, where local SDR devices perform primary processing, signature encoding and classification in real time, while the cloud environment provides centralised storage of the global signature database, inter-device synchronisation of updates, and execution of resource-intensive procedures for retraining and analysis of undefined cases. This division made it possible to clearly show at which point the state vector S transitions from "present channel description" to "signature description of the impact", how the suspicious fragment is isolated and the descriptor is formed, and where exactly the decision is made between classification based on known patterns and initiation of incremental replenishment of the knowledge base in the case of an unknown signature. This coordinated sequence of blocks and branches, as well as the transition points between the "continue work" and "in-depth analysis → comparison → decision/update" modes, are shown in Figure 1.

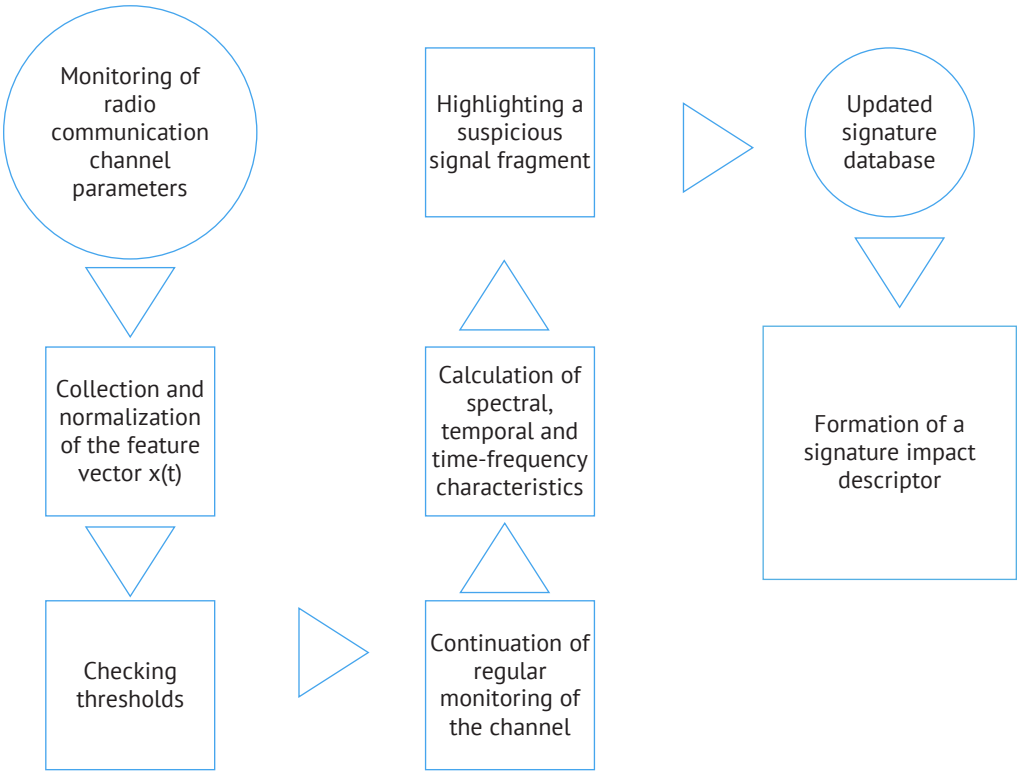


Figure 1. Block diagram of an algorithmic conveyor for determining the type of intentional interference
Source: compiled by the authors using scikit-learn and Python 3.12

Figure 1 shows that the input object of the conveyor is the channel state vector , formed in the communication parameter monitoring mode, after which the event detection threshold block is launched (control of critical deviations, in particular for BER). If the threshold conditions are not met, the system switches to the mode of continuing operation without activating extended processing, which, in the logic of the conveyor, means that there are no grounds for a signature description as a fault event. If the detection criteria are met, the pipeline moves to the advanced processing branch, where the suspicious fragment is selected, the spectral and temporal characteristics are calculated, and the interference descriptor is formed as a compact signature representation. The next central stage is a metric comparison of the formed signature

with the knowledge base templates: if there is a match, a classification decision is formed, and if there is no match, a mechanism for creating a new template and adding a new signature to the base is activated, which implements incremental knowledge expansion without complete retraining.

Thus, the key result of algorithmic formalisation is recorded: the method is presented as a reproducible pipeline with clearly defined mandatory blocks and unambiguous interfaces between them, which eliminates the variability of interpretation of "when exactly" and "which exactly" features are transferred to the signature, as well as at what point the basis for the classification decision arises. In this logic, the detection node acts as a flow control gateway: it filters background parameter fluctuations and initiates resource-intensive operations only when there is a consistent degradation of the channel across several indicators, i.e., it minimises unnecessary calculations and reduces the risk of "reclassifying" normal modes as interference. Segmentation and feature calculation form stable quasi-stationary fragments, within which the signature stage provides not just a description of the signal, but a compact and comparable vector representation of the impact: it unifies energy, error, spectral, and structural manifestations in a common space suitable for metric comparison. The final pair of nodes, "comparison → decision/update", sets the dual-loop behaviour of the system: for known classes, it supports fast pattern recognition based on maximum similarity, and for undefined cases, it introduces a formal rejection mechanism with a threshold S, which prevents forced assignment to the false class and transfers such observations to a controlled accumulation mode. This combination of "classification + reject option + incremental replenishment" renders the conveyor suitable for operation in conditions of incomplete a priori information: the system not only recognises what is already described in the database, but also has a formalised way of expanding the signature database without destroying the decision-making logic and without the need to completely reconfigure each stage.

Results of classification and rejection of unknown signature profiles

Within the framework of generalising the results of algorithmic implementation, formalised decision-making rules have been established that distinguish between assigning a signature to a known class and controlled rejection of undefined cases. The classification node of the method is presented as a threshold-metric procedure, in which the decision is based not only on the maximum similarity to the template, but also on checking the sufficiency of this similarity relative to the established membership threshold. This approach makes it possible to avoid the forced assignment of any observation to an existing class and to introduce a separate mode for processing new or atypical signatures. The generalised resulting rules for comparing metrics, threshold criteria and decision types are given in Table 3.

Table 3. Resultant classification rules and signature rejection thresholds (reject option)

Metric/assessment of belonging	Settlement terms	Threshold criterion	Type of solution	System operation	Status of entry in the signature database
Cosine similarity	$\max_i \cos(x, s_i)$	$\geqslant \tau\_sim$	Known class	Assigning a class with the maximum similarity value	Existing template is used
Cosine similarity	$\max_i \cos(x, s_i)$	$< \tau\_sim$	Rejection	The signature is marked as unknown	Temporary candidate record
Euclidean distance	$\min_i d(x, s_i)$	$\leqslant \tau\_dist$	Known class	Assignment of nearest class	Existing template is used
Euclidean distance	$\min_i d(x, s_i)$	$> \tau\_dist$	Rejection	Switching to accumulation mode	Temporary candidate record
Probabilistic estimation of the classifier	$\max_k P(y_k \mid x)$	$\geqslant \tau\_prob$	Known class	Classification decision by argmax	Existing template is used
Probabilistic estimation of the classifier	$\max_k P(y_k \mid x)$	$< \tau\_prob$	Reject option	Refusal of compulsory classification	New template after confirmations
Agreed multimetric assessment	$(sim \geqslant \tau\_sim) \wedge (P \geqslant \tau\_prob)$	Both thresholds were met	Confirmed class	Decisions with increased confidence	Updating template statistics
Agreed multimetric assessment	Thresholds are not aligned	Partial fulfilment	Undefined case	Re-examination/ additional signs	Deferred candidate

Source: compiled by the authors using scikit-learn and MATLAB R2024a

Analysis of the generalised rules in Table 3 showed that the effective decision-making logic in the method reduces to a formalised division of two modes: rapid typing according to known patterns and controlled "non-decision-making" in cases where the signature does not have sufficient similarity to any of the

existing classes. For the metric approach, this is implemented symmetrically: for cosine similarity, "membership" is determined by exceeding the threshold τ_sim for the maximum similarity, while for Euclidean distance, it is determined by not exceeding τ_dist for the minimum distance; in both cases, the decision about the known class is the result of not a single comparison, but competition with the entire set of templates, which makes the classification dependent on the relative position of the signature in the feature space. At the same time, the reject option mechanism introduces a fundamentally different condition: if none of the templates meets the specified proximity criterion, the signature is not "forcibly" assigned to the nearest class, but is transferred to the rejection mode, i.e., it becomes an object for accumulation and further incremental replenishment of the knowledge base. This branch minimises the risk of misclassification in conditions of uncertain interference, when the input influence may be new or combined and formally closer to one of the classes only due to random fluctuations in features.

The threshold mechanism can be implemented not only through "geometric" metrics (cos, d), but also through a probabilistic interpretation of the classifier output, where the criterion becomes the maximum a posteriori estimate $\max_k P(y_k|x)$ relative to τ_prob. This unified reject option for different types of models: regardless of whether the decision is obtained as a comparison with benchmarks or as the output of a multi-class classifier, rejection is formed according to the same logic of "insufficient confidence → unknown type". Another crucial result is the introduction of a consistent rule for multimetric verification: class confirmation is considered most robust when both the similarity threshold and the probability threshold are met simultaneously, rather than just one of them. Instead, partial fulfilment of thresholds (e.g., high similarity without sufficient P or vice versa) is interpreted as an uncertain case and transferred to a re-verification mode or descriptor completion mode, which reduces the sensitivity of the system to "borderline" signatures between classes. As a result, the resulting decision-making logic is recorded as a two-level procedure "maximum similarity → threshold verification → class or rejection", which simultaneously ensures stable recognition of known types of interference and a formalised mechanism for working with new signatures. This scheme increases the system's resistance to uncertainty in the interference environment and supports the controlled expansion of the signature database without compromising the reproducibility of the algorithmic pipeline.

Model-simulation validation of the method in streaming SDR scenarios

In the note on the reproducibility test results of the proposed method, its algorithmic pipeline was integrated into a stream-oriented SDR signal processing architecture and tested in a series of controlled parameterised simulation scenarios. Validation was performed not at the level of individual features or an isolated classifier, but at the level of the complete working chain "signal generation → channel → noise superimposition → feature coding → signature comparison → decision", which corresponds to the logic of the computational pipeline adopted in the methodology. This approach verified not only the correctness of classification decisions but also the consistency of interfaces between blocks, stability of operation in streaming mode, and repeatability of results at fixed scenario parameters. A generalised structural diagram of the simulation test is shown in Figure 2.

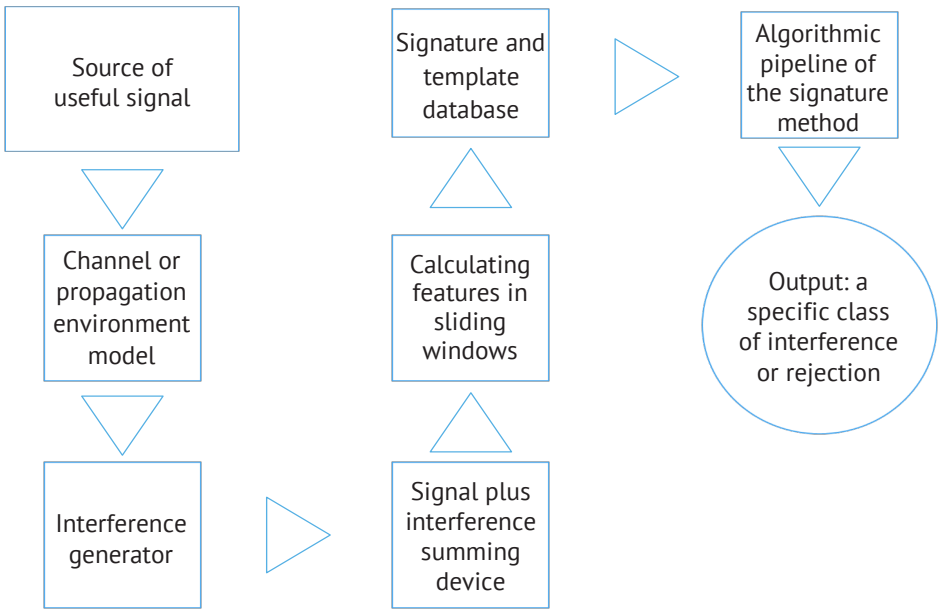


Figure 2. Generalised diagram of SDR-oriented simulation of method flow checking

Source: compiled by the authors using GNU Radio 3.10.0.0 and Python 3.12

The diagram in Figure 2 shows that the validation circuit consists of a series of functional blocks for signal synthesis and processing, in which a useful signal source with controllable modulation parameters is combined with a parametric generator of interference influences. The interference block specifies the type of interference, intensity, spectral width, time structure and duration, which can be used to form reproducible signature profiles for impulse, noise, harmonic, modulation and structured low-visibility interference. After the signal and interference are summed, the stream enters the feature analysis block, where SNR, RSSI, BER indicators, PSD, and time-frequency descriptors are calculated in sliding windows, consistent with the formalised channel state vector. Next stage implements the embedding of the algorithmic conveyor method into stream processing: the detection node performs threshold control of parameter degradation and controls the activation of extended signature analysis, the segmentation block selects quasi-stationary fragments, after which a compact signature representation is formed, and a metric comparison with the template database is performed. The classification module and threshold rejection mechanism operate in the same mode as in the formalised algorithmic scheme, ensuring the identity of the decision logic between the "offline" description of the method and its stream implementation.

The repeatability of scenarios was ensured by fixing the parameters of signal and noise generators, sampling frequency, window lengths, spectral estimation parameters, and detection and rejection threshold values. With unchanged configuration parameters, simulation runs reproduced the same signature profiles and the same classification decisions, confirming the invariance of the method to the order of block processing and its suitability for reproducible software implementation. Additionally, the study established that the division of the conveyor into a light monitoring circuit and a conditionally activated signature circuit in a streaming SDR scheme retains its effectiveness: resource-intensive feature calculations and metric comparisons are performed only when event triggers are present, which is consistent with the requirements for limited delays and computational resources of SDR systems. Simulation testing confirmed the method's suitability for integration into the SDR architecture: the algorithmic pipeline maintains structural integrity, feature and signature interfaces remain consistent, and the classification and rejection logic works reproducibly in controllable parameterised scenarios, confirming the practical feasibility of the proposed scheme.

Computational complexity and resource efficiency of the SDR classification algorithm

As part of summarising the results of the algorithmic implementation, an analytical decomposition of computational costs by pipeline stages was performed to verify the suitability of the method for SDR-oriented systems with limited resources. The result obtained presents complexity not as a single estimate, but as the sum of the costs of individual functional blocks – from feature formation to metric comparison and decision making. This decomposition establishes which stages determine the upper limit of processing time and how complexity scales with the growth of the feature vector dimension and the size of the signature database. Generalised estimates of complexity and resource requirements by stage are presented in Table 4.

Table 4. Assessment of the computational complexity of the algorithm stages and the $O(n \cdot m)$ dependency

Conveyor stage	Base operations	Asymptotic complexity	Dependent parameters	Resource profile
Window formation of features	Updating statistics, aggregates in a sliding window	O(n)	n – number of signs	Linear, streaming
Spectral and time-frequency estimates	PSD, local spectra	O(n log n)	n – window size	Moderate, block-based
Event detection	Threshold checks	O(n)	n – signs	Low
Segmentation	Adaptive window, stability	O(n)	n – signs	Low
Dimensionality reduction	Projection/auto encoder	O(n·k)	k – latent dimension	Controlled
Metric comparison with the base	Cos/distance across all templates	O(n·m)	n – dimension, m – templates	Dominant stage
Threshold decision	Comparison with	O(m)	m – templates	Low
Base update	Record addition	O(n)	n – signs	Episodic

Source: compiled by the authors using analytical and programmatic verification in Python 3.12

Analysis of generalised estimates in Table 4 showed that the computational profile of the method has a pronounced modular structure, in which most stages are characterised by linear complexity in terms of the dimension of the feature vector. The calculation of statistical, energy, and time descriptors in sliding windows is implemented as a streaming procedure with linear updating of aggregates, which does not require a complete recalculation at each step. Spectral estimates have a logarithmic factor, but are performed only for segments that have passed the detection filter and therefore do not form a constant load in normal

monitoring mode. The cost-determining stage is the metric comparison of the signature with the template database, for which the order O(n · m) is established. This dependence directly follows from the need to calculate the measure of proximity between the input vector of dimension n and each of the m signature templates. At the same time, the increase in complexity is linear in both the dimension of the features and the size of the database, without quadratic or combinatorial terms, which preserves the predictability of the response time when incrementally expanding the signature repository. Reducing the dimension before metric comparison further reduces the effective n and thus directly reduces the cost of this stage. Analysis of the dependence of the classification decision delay on the size of the signature database (Fig. 3) confirmed the linear nature of the processing time scaling in accordance with the theoretically established complexity O(n · m).

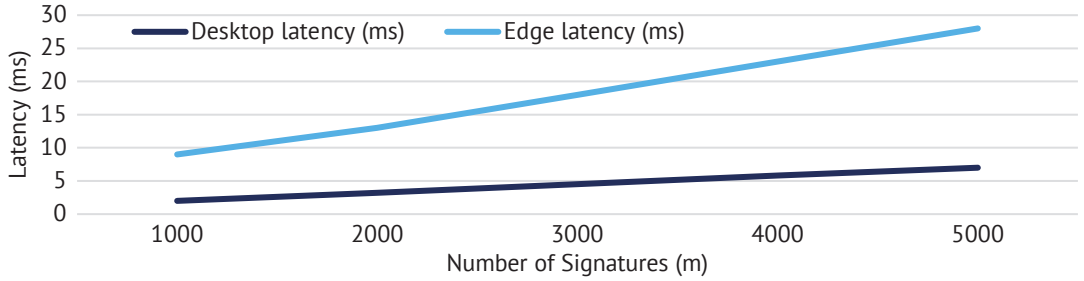


Figure 3. Dependence of classification decision delay on signature database size (m) in streaming SDR mode for desktop and edge configurations

Source: compiled by the authors using GNU Radio 3.10.0.0 and Python 3.12

Following Figure 3, with a fixed latent signature dimension k = 16 and an initial feature vector dimension n = 64, the average event cycle delay in the desktop configuration increased from ≈2 ms at m = 1,000 to ≈7 ms at m = 5,000, while in the edge configuration, it increased from ≈8-10 ms to ≈25-30 ms in the same range. The resulting curves have a near-linear slope without sharp bends, which indicates the absence of hidden quadratic or combinatorial components in the implementation and confirms the dominance of the metric comparison stage of the signature with the template base. At the same time, in normal monitoring mode (without activating the signature branch), the delay remained at the sub-millisecond level, which ensures the suitability of the pipeline for streaming SDR scenarios with strict real-time response requirements. The memory profile assessment showed a linear dependence of the amount of RAM on the number of stored templates (Fig. 4).

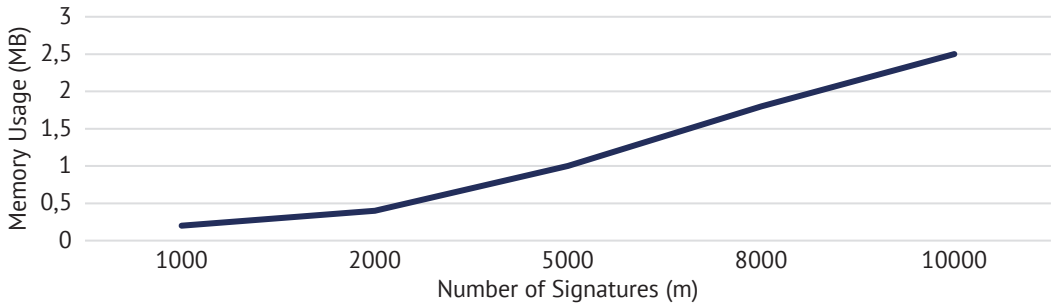


Figure 4. Dependence of the amount of RAM required to store the signature database on the number of templates (m) with latent dimension k = 16

Source: compiled by the authors

Following Figure 4, for signatures in float32 format and latent dimension k = 16, the memory cost of a single record, including service metadata, averaged 0.15-0.25 KB, resulting in approximately 0.2 MB at m = 1,000, 1 MB at m = 5,000, and up to 2-2.5 MB at m = 10,000. Even considering sliding window buffers, spectral estimates, and auxiliary data structures, the total RAM consumption in the edge configuration remained within 10-20 MB, which corresponds to the typical resource limitations of embedded SDR platforms. Thus, scaling the signature database does not lead to a jump in memory consumption and remains predictable and controllable. The set of results obtained confirmed that the software and system implementation of the method is characterised by a dual-loop resource profile: low constant latency in monitoring mode and linearly scalable latency and memory consumption in event mode. This ensures the practical applicability of the algorithmic pipeline for implementation on SDR platforms with limited resources, provided that the signature base grows in a controlled manner and latent feature coding is used. Intentional radio interference

can be reproducibly identified in SDR systems through a coordinated multidimensional representation of the channel state: heterogeneous parameters (SNR, RSSI, BER, PSD, and time/time-frequency descriptors) are reduced to a normalised vector x(t), based on which signature profiles of five basic classes of interference are formed as stable multi-parameter patterns. The method is implemented as an integrated computational pipeline with two operating modes (monitoring and conditionally activated extended processing), where decisions are made through metric/probabilistic comparison with the template database and supplemented by a formal reject option mechanism for controlled rejection and incremental replenishment of the database with new signatures. Stream-oriented SDR simulation testing showed reproducibility of interfaces between blocks and decisions for fixed scenario parameters, and an analytical complexity assessment established a controllable resource profile with a dominant O(n·m) matching stage, suitable for low-resource/edge-SDR implementation due to the detection "gateway" and reduction of signature dimensions.

DISCUSSION

The results obtained showed that the formalisation of the multidimensional feature space of the channel, the signature representation of influences, and the threshold-metric decision-making procedure are consistent with current approaches in which the separation of interference types is achieved through time-frequency, spectral, and structural representations of the signal. The study demonstrated that the decisive role of the time-frequency structure for separating interference classes was consistent with the approach of P. Wang *et al.* (2022), where component-oriented time-frequency representation together with convolutional neural network processing was used for wireless interference classification, and class separation was associated with stable local time-frequency components. This formulation is consistent with the interpretation using spectral and time-frequency descriptors as a mandatory block of the signature profile in the proposed method, but differs in that signature matching is fixed as a formalised metric procedure with explicit rules for normalisation and window matching.

Further development of intelligent approaches to detecting and suppressing interference in software-defined radio systems is reflected in the results of S. Nanayakkara *et al.* (2025), where algorithmic solutions for jamming detection and modulation recognition were combined with software-configurable radio architecture and machine learning models. The study demonstrated that the effectiveness of such systems is determined by the consistency between feature extraction and classification blocks, which directly correlates with the conveyor scheme introduced in the method: "feature coding → signature descriptor → metric comparison → decision", in which the interfaces between stages are fixed and reproducible. A similar emphasis on time-frequency analysis as the basis for deep classification of interference was reflected in the results of C. Xu *et al.* (2022), where for systems with stepwise frequency restructuring, the deep recognition model was built on time-frequency features and demonstrated increased resolution of complex interference effects. This pattern corresponded to the role established in the results of time-frequency and entropy descriptors as key indicators of structured and subtle influences, particularly in scenarios with low energy contrast, where simple energy indicators do not provide reliable typification.

The need to combine the recognition of known attacks with the formalised detection of unknown influences was consistent with the approach of S. Hong *et al.* (2023), where a hybrid detection algorithm provided simultaneous classification of known types of jamming and a separate mode for detecting unknown attacks. It was shown that such dual-mode logic was consistent with the threshold mechanism for signature rejection introduced in the proposed method, whereby insufficient similarity to templates does not lead to forced classification, but transfers the observation to an undefined profile mode with subsequent incremental replenishment of the database. Further comparison of the results with current research on monitoring and countering radio jamming showed that the multi-parameter approach to describing the channel state and the signature typology of interference are consistent with the approaches in which interference is considered as a reproducible multidimensional profile of changes in indicators, rather than as an isolated anomaly of a single parameter. This approach was consistent with the review results of Y. He *et al.* (2024), where monitoring of continuous noise in navigation systems was interpreted as a multi-level system of indicators with simultaneous consideration of energy, spectral and structural manifestations, which, according to the interpretation, directly correlates with the formalisation of the channel state vector and block grouping of features in the signature space.

The focus on computationally efficient intelligent algorithms for detecting jamming of embedded and unmanned radio systems was reflected in the results of S. Cibecchini *et al.* (2025), which showed that lightweight machine learning models provide practical applicability in resource-constrained environments, provided that prior feature compression is performed. This conclusion is consistent with the use of signature dimension reduction and localisation of resource-intensive procedures only in a conditionally activated branch after event detection in the proposed method, which ensures a controllable computational profile. A similar emphasis on stream detection with window processing and hybrid real-time models was reflected in the results of J. Sormayli *et al.* (2025), where jamming detection was performed in sliding time windows with early warning indicators before receiver saturation. This logic was consistent with the detection node and

segmentation block in the method pipeline, where window matching and threshold triggers determine the moment of transition to extended signature analysis. The development of deep neural network methods for anti-jamming processing in navigation receivers was reflected in the results of Y. Xie *et al.* (2025), which showed that a deep neural network in the antenna array of a navigation receiver provides improved noise immunity through trained spatial-spectral filtering. This pattern is consistent with the interpretation of the classification block as one of the possible decision-making tools in the signature pipeline, but not the only criterion, since in the proposed scheme, the decision is additionally limited by metric and threshold membership conditions.

The need for open recognition of interference signals with the separation of unknown classes was reflected in the results of X. Chen *et al.* (2022), where effective recognition in an open set of classes was achieved through prototype-oriented convolutional coding and threshold checking of the distance to class representatives. This approach is conceptually consistent with the mechanism of threshold rejection of signatures and incremental formation of new patterns without forced classification introduced in the method. Use of machine learning to mitigate the impact of radio jamming at the physical level of communication was consistent with the results of M. Jacovic *et al.* (2023), which showed that training models can be integrated directly into physical-level anti-jamming procedures. According to the interpretation, this corresponds to the integration of a probabilistic classifier into a formalised signature-metric scheme, where the training model operates as one of the nodes, but within the limits of a reproducible algorithmic pipeline. The consideration of signal identification and interference as a multi-task spectral sensing problem was reflected in the results of H. Kim *et al.* (2023), where multi-task deep learning was used to simultaneously identify signal types and spectral states. This approach was consistent with the multi-block feature structure in the signature method, where energy, error, spectral, and time-structure descriptors are considered as coordinated but functionally different components of a single state vector.

A comparison of the results obtained with publications covering navigation systems, wireless networks, aviation communication channels, and integrated space-air-ground networks showed that the proposed signature approach with a formalised channel state vector, window matching, and threshold rejection of unknown profiles is consistent with the current trend of transition from single indicators to multi-parameter recognition schemes, where the critical factor is not the "strength" of a single feature, but the reproducible profile of coordinated changes in time, frequency, and time-frequency domain. The generalised state of development of methods for detecting jamming and substitution of navigation signals is reflected in the results of K. Radoš *et al.* (2024), which show that effective procedures for detecting intentional interference require a combination of energy, correlation, spectral and time indicators, as well as the separation of the tasks of "event detection" and "impact type identification", which conceptually corresponds to the introduction of two conveyor operating modes in the method and the structural division of features into functional blocks. The practical focus on low-resource implementations and hardware-accessible measurements in navigation receivers was emphasised by S. Kocher *et al.* (2025), where interference classification was performed at the peripheral computing level using inexpensive commercial off-the-shelf equipment, and the robustness of the solution was linked to the consistency of the model and features under limited resources; this logic directly correlates with the conclusion obtained in the results about the dominance of the metric comparison stage and the need to control the dimension of signatures and the size of the library for the expected delay. A review by K.J. Silva Lorraine & M. Ramarakula (2021) confirmed that neural network approaches to countering jamming in navigation systems most often rely on time-frequency or spectral representations, however, their reliability in applied scenarios is determined by the stability of the feature representation and the availability of procedures that prevent the forced classification of non-standard influences; in this sense, the mechanism of threshold rejection and controlled accumulation of unknown profiles introduced in the method forms precisely the link that overview generalisations often identify as problematic for "closed" settings. The study by A. Ghanbarzade & H. Soleimani (2025) further emphasised the diversity of machine and neural network models for identifying substitution and jamming in navigation systems, while also emphasising that the suitability of such models is determined by the formulation of the problem and the correctness of the membership criteria, which is consistent with the given formalisation of decision rules through similarity thresholds and probabilistic membership and the interpretation of the "unknown" as a separate mode rather than a classification error.

At low-power wireless network level, E. Jayabalan & R. Pugazendi (2022) demonstrated the effectiveness of deep models for detecting jamming attacks, but from an engineering point of view, the key point is that the effectiveness of such approaches depends on correct window coding and the consistency of features with the channel mode; This directly correlates with the conclusion that fixed normalisation rules and window alignment are critical for metric comparability of signatures between scenarios. The transition from local interference classification to the broader concept of "cognitive" recognition of the electromagnetic spectrum state is consistent with the findings of H. Zhao *et al.* (2025), where intelligent countermeasures to jamming are interpreted as a technology based on spectrum feature recognition and decision adaptation; in the formulation,

this is reflected in the consideration of signatures as an operational "language" for a knowledge base that can be incrementally replenished without destroying the logic of the decision. The practical aviation context of "manned-unmanned" real-time communication channels in the study by R.D. Jurado *et al.* (2025) showed that the integration of interference detection modules into onboard circuits requires not only accuracy but also reproducible interfaces between processing stages and controlled delays, which functionally coincides with the presented pipeline formalisation and decomposition of complexity into blocks.

The extension of the problem to the level of "electromagnetic environment" for integrated "space-air-ground" networks is reflected in the study by J. An *et al.* (2026), where situational awareness and spectral environment modelling are interpreted as a systemic function of the network; in terms of content, this is consistent with the formulation in which the channel state vector and signature library act as a local mechanism for situational recognition of types of influences with the possibility of expansion. The task of open recognition of combined interference without forced typification is directly developed in the study by Y. Xiao *et al.* (2024), where the open mode of recognition of composite jamming signals is implemented through multitasking and multi-target learning; conceptually, this correlates with the conclusion about the need to reject unknown profiles and the controlled accumulation of candidates for new signatures, especially for combined influences that do not match any template. The focus on synergy between machine learning and classical signal processing in low-resource environments is systematically reflected in J.R. van der Merwe *et al.* (2024), where it is shown that optimal solutions for interference classification under limited resources are achieved not by "maximising model complexity" but by the right combination of features, pre-processing and compact classifiers; This practically confirms the decomposition of complexity shown and the advisability of reducing the dimension before metric comparison as a direct way to control the dominant cost component. In summary, comparison with relevant studies has shown that the results obtained support the current trend of transition from single indicators to multi-parameter time-frequency and spectral-structural representations. At the same time, the proposed approach differs from predominantly "model-oriented" solutions in that it fixes a reproducible pipeline with explicit rules for normalisation, window matching and threshold rejection, which simultaneously ensures stable typification of known classes and controlled work with unknown profiles through the accumulation and incremental expansion of the signature library for a predicted resource-intensive profile.

CONCLUSIONS

Within the scope of the study, a reproducible algorithm-oriented procedure for determining types of intentional radio interference for software-controlled radio systems has been developed based on a coordinated multidimensional representation of the channel state, signature typology, and threshold-metric decision-making logic with a mechanism for rejecting uncertain cases. It is shown that heterogeneous channel observation parameters (SNR, RSSI, BER, PSD) in combination with time statistics and time-frequency descriptors can be reduced to a normalised state vector x(t) suitable for metric comparison; the key condition for the correctness of the comparison is fixed rules of normalisation and window matching, which reduce the influence of absolute levels and random fluctuations and make the similarity of signatures dependent on the type of interference. Based on this coordinated feature space, signature profiles of five basic classes of interference (pulse, broadband noise, harmonic, modulation, and low-perceptible LPI/LPD) are formed, and the study established that the separation of classes is ensured not by the extremity of a single indicator, but by stable multi-parameter combinations of energy, error, spectral and time-structural manifestations in sliding windows, sufficient for template comparison and scalable replenishment of the database.

The algorithm is recorded as an integral computational pipeline with unambiguous interfaces: "state vector → event detection → segmentation → signature formation → (if necessary) dimensionality reduction → metric/probabilistic comparison → decision", where the detection node acts as a control gateway and activates resource-intensive operations only when there is an agreed degradation of channel parameters. The introduction of a reject option with thresholds τ ensured the formal separation of the "known class" and "unknown signature", preventing forced typing and transferring atypical observations to a controlled mode of accumulation and confirmation before incremental addition to the database. Streaming SDR-oriented simulation verification confirmed the reproducibility of solutions at the full processing chain level for fixed scenario parameters, and analytical cost estimation established a controllable resource profile with a dominant $O(n \cdot m)$ comparison and the possibility of reducing the effective n through latent coding, making the approach suitable for low-resource/edge-SDR implementations. Further research should address calibration of thresholds τ on broader sets of scenarios and experimentally validating signatures in real SDR paths under combined and adaptive interference.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] An, J., Tao, Y., Zhang, X., Hua, Z., Wang, S., Pan, G., Yang, X., Niyato, D., & Karagiannidis, G.K. (2026). Electromagnetic situation awareness and modeling for space-air-ground integrated networks. *National Science Review*, 13(1), article number nwaf492. doi: 10.1093/nsr/nwaf492.

[2] Boholii, S., Hurskyi, T., Makarchuk, V., & Khyzhyi, A. (2022). Increasing the anti-jammingness of mobile radio networks with adaptive beamforming. *Systems and Technologies of Communication, Informatization and Cyber Security*, 2(2), 5-14. doi: 10.58254/viti.2.2022.01.5.

[3] Chen, X., Zhao, Z., Ye, X., Zheng, S., Lou, C., & Yang, X. (2022). Efficient open-set recognition for interference signals based on convolutional prototype learning. *Applied Sciences*, 12(9), article number 4380. doi: 10.3390/app12094380.

[4] Cibecchini, S., Chiti, F., & Pierucci, L. (2025). A lightweight AI-based approach for drone jamming detection. *Future Internet*, 17(1), article number 14. doi: 10.3390/fi17010014.

[5] Davies, N., Dogan, H., & Ki-Aries, D. (2025). Application of systems-of-systems theory to electromagnetic warfare intentional electromagnetic interference risk assessment. *Systems*, 13(4), article number 244. doi: 10.3390/systems13040244.

[6] Ghanbarzade, A., & Soleimani, H. (2025). GNSS/GPS spoofing and jamming identification using machine learning and deep learning. *ArXiv*. doi: 10.48550/arXiv.2501.02352.

[7] Han, H., Li, W., Feng, Z., Fang, G., Xu, Y., & Xu, Y. (2022). Proceed from known to unknown: Jamming pattern recognition under open-set setting. *IEEE Wireless Communications Letters*, 11(4), 693-697. doi: 10.1109/LWC.2021.3140145.

[8] He, Y., Li, B., Chen, J., Wang, Z., Xiao, W., & Lu, Z. (2024). Overview of the development of satellite navigation blanket interference monitoring. *Frontiers in Physics*, 12, article number 1487384. doi: 10.3389/fphy.2024.1487384.

[9] Hong, S., Kim, K., & Lee, S.H. (2023). A hybrid jamming detection algorithm for wireless communications: Simultaneous classification of known attacks and detection of unknown attacks. *IEEE Communications Letters*, 27(7), 1769-1773. doi: 10.1109/LCOMM.2023.3275694.

[10] Hurskyi, T., Panchenko, I., Voskolovych, O., Saliy, O., & Kotenko, I. (2025). Evaluation of the efficiency of radio electronic suppression of FPV UAV control channels. *Systems and Technologies of Communication, Informatization and Cyber Security*, 8, 34-44. doi: 10.58254/viti.8.2025.03.34.

[11] Jacovic, M., Rey, X.R., Mainland, G., & Dandekar, K.R. (2023). Mitigating RF jamming attacks at the physical layer with machine learning. *IET Communications*, 17, 12-28. doi: 10.1049/cmu2.12461.

[12] Jayabalan, E., & Pugazendi, R. (2022). Deep learning model-based detection of jamming attacks in low-power and lossy wireless networks. *Soft Computing*, 26, 12893-12914. doi: 10.1007/s00500-021-06111-7.

[13] Jurado, R.D., Reina, B.J., Suárez, M.Z., Moreno, F.P., Arnaldo, C.G., & Valdés, R.M. (2025). AI-driven real-time interference detection in manned-unmanned aircraft communications: Concept of operations and integration. *Aerospace Science and Technology*, 168, article number 110722. doi: 10.1016/j.ast.2025.110722.

[14] Kantsedal, V., & Mogyla, A. (2025). Features of constructing an algorithm for the cycle between stage-by-stage situational control of conflict in-teraction of the ground-based RES complex with small (light) drones. *Radiotekhnika*, 221, 89-106. doi: 10.30837/rt.2025.2.221.12.

[15] Kim, H., Kim, Y.J., & Kim, W.T. (2023). Multitask learning-based deep signal identification for advanced Spectrum sensing. *Sensors*, 23(24), article number 9806. doi: 10.3390/s23249806.

[16] Kocher, S., Contreras Franco, D., Dietz, A., & Rügamer, A. (2025). On the edge model-aided machine learning GNSS interference classification with low-cost COTS hardware. *Engineering Proceedings*, 88(1), article number 51. doi: 10.3390/engproc2025088051.

[17] Krayani, A., Alam, A. S., Marcenaro, L., Nallanathan, A., & Regazzoni, C. (2022). Automatic jamming signal classification in cognitive UAV radios. *IEEE Transactions on Vehicular Technology*, 71(12), 12972-12988. doi: 10.1109/TVT.2022.3199038.

[18] Li, Y., & Pawlak, M. (2022). Jamming detection and classification in OFDM-based UAVs via feature- and spectrogram-tailored machine learning. *IEEE Access*, 10, 16859-16870. doi: 10.1109/ACCESS.2022.3150020.

[19] Nanayakkara, S., Sumathipala, S., Karunanayake, N., Karunanayake, M., & Kumara, T. (2025). Smart drone neutralization: AI driven RF jamming and modulation detection with software defined radio. *Drones and Autonomous Vehicles*, 2(4), article number 10019. doi: 10.70322/dav.2025.10019.

[20] Radoš, K., Brkić, M., & Begušić, D. (2024). Recent advances on jamming and spoofing detection in GNSS. *Sensors*, 24(13), article number 4210. doi: 10.3390/s24134210.

[21] Rijnsdorp, J., van Zwol, A., & Snijders, M. (2023). Satellite navigation signal interference detection and machine learning-based classification techniques towards product implementation. *Engineering Proceedings*, 54(1), article number 60. doi: 10.3390/ENC2023-15449.

[22] Silva Lorraine, K.J., & Ramarakula, M. (2021). A comprehensive survey on GNSS interferences and the application of neural networks for anti-jamming. *IETE Journal of Research*, 69(7), 4286-4305. doi: 10.1080/03772063.2021.1953407.

[23] Sormayli, J., Darvishi, M., Zarrinnegar, K., & Mosavi, M.R. (2025). Real-Time jamming detection using windowing and hybrid machine learning models for pre-saturation alerts. *Scientific Reports*, 15, article number 24748. doi: 10.1038/s41598-025-10567-0.
[24] van der Merwe, J.R., Franco, D.C., Feigl, T., & Rügamer, A. (2024). Optimal machine learning and signal processing synergies for low-resource GNSS interference classification. *IEEE Transactions on Aerospace and Electronic Systems*, 60(3), 2705-2721. doi: 10.1109/TAES.2023.3349360.
[25] Wang, P., Cheng, Y., Dong, B., Peng, Q., & Li, S. (2022). Multi-domain networks for wireless interference recognition. *IEEE Transactions on Vehicular Technology*, 71(6), 6534-6547. doi: 10.1109/TVT.2022.3164908.
[26] Wang, P., Cheng, Y., Shang, G., Wang, J., & Li, S. (2022). Time-frequency component-aware convolutional neural network for wireless interference classification. *IEEE Wireless Communications Letters*, 11(12), 2487-2491. doi: 10.1109/LWC.2022.3204756.
[27] Xiao, Y., Zhang, R., Yu, X., & Jiang, Y. (2024). Open-set recognition of compound jamming signal based on multi-task multi-label learning. *IET Radar, Sonar & Navigation*, 18(8), 1235-1246. doi: 10.1049/rsn2.12561.
[28] Xie, Y., Chen, L., Zhang, K., Huang, X., Peng, J., & Ni, S. (2025). Deep neural network based anti-jamming processing in the GNSS array receiver. *GPS Solutions*, 29, article number 63. doi: 10.1007/s10291-025-01821-z.
[29] Xu, C., Ren, J., Yu, W., Jin, Y., Cao, Z., Wu, X., & Jiang, W. (2022). Time-frequency analysis-based deep interference classification for frequency hopping system. *EURASIP Journal on Advances in Signal Processing*, 2022, article number 90. doi: 10.1186/s13634-022-00913-z.
[30] Zhao, H., Zhao, G., Wang, X., Zhang, Z., & Xun, X. (2025). Intelligent anti-jamming communication technology with electromagnetic spectrum feature cognition. *PloS One*, 20(4), article number e0319953. doi: 10.1371/journal.pone.0319953.
[31] Zylevich, V., & Svakha, S. (2025). Highly efficient methods of determining radio interference based on communication signal analysis. *Scientific Notes of V.I. Vernadsky Ternopil National University. Series: Technical Sciences*, 36(75), 78-83. doi: 10.32782/2663-5941/2025.1.2/12.

Метод визначення типу навмисних завад в умовах невизначеності завадової обстановки на основі відомих моделей

Григорій Радзівілов

Кандидат технічних наук, доцент
Військовий інститут телекомунікацій та інформатизації імені Героїв Крут
01011, вул. Князів Острозьких, 45/1, м. Київ, Україна
https://orcid.org/0000-0002-6047-1897

Дмитро Павлюк

Ад'юнкт
Військовий інститут телекомунікацій та інформатизації імені Героїв Крут
01011, вул. Князів Острозьких, 45/1, м. Київ, Україна
https://orcid.org/0000-0001-8461-3899

Анотація. Метою дослідження було теоретично обґрунтувати метод ідентифікації типу навмисних завад у невизначеній завадовій обстановці для програмно-визначуваних радіосистем шляхом використання відомих моделей завад і формалізованого аналізу вектора стану каналу зв'язку. Методологічна база включила віконне обчислення узгодженого вектора ознак, їх нормалізацію, сигнатурне кодування, метрикне та ймовірнісне порівняння з бібліотекою шаблонів, порогове відторгнення невизначених випадків і модельно-симуляційну перевірку в потоковій архітектурі програмно-визначуваного радіо. У результаті сформовано формалізований ознаковий простір і сигнатурні профілі п'яти базових класів завад – імпульсних, широкосмугових шумових, гармонічних, модуляційних і малопомітних – як стійкі багатопараметричні патерни спільної зміни показників сигналу, що забезпечують роздільність класів у різних режимах співвідношення сигналу до шуму, варіативності ймовірності бітової помилки та морфології спектральної густини потужності і відтворюються при повторних прогонах за незмінних параметрів. Метод реалізовано у вигляді обчислювального конвеєра з послідовністю етапів «вектор стану → детекція події → сегментація → сигнатурний опис → зниження розмірності → порівняння → рішення», де ресурсоємні операції активуються лише за наявності подієвого тригера, а інтерфейси між блоками фіксують однозначний перехід від поточних спостережень до компактної сигнатури впливу. Сформульовано результатні правила класифікації з порогами належності та процедурою керованого відторгнення, що переводить нові сигнатури в режим накопичення та інкрементального доповнення бази із подальшим підтвердженням стабільності профілю на повторних спостереженнях. Симуляційна перевірка на серії параметризованих сценаріїв показала відтворюваність сигнатурних профілів і класифікаційних рішень за фіксованих налаштувань. Практична значущість результатів полягає у можливості впровадження методу розробниками та інтеграторами вбудованих програмно-керованих радіосистем і систем радіомоніторингу для потокового визначення типу навмисної завади за сигнатурними профілями з пороговим відторгненням невідомих випадків та керованим поповненням бібліотеки сигнатур

Ключові слова: вектор стану каналу; часово-частотні дескриптори; обчислювальний конвеєр; сигнатурне кодування; reject option

BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 2. 2026

UDC 004.946:004.921
DOI: 10.62660/bcstu/2.2026.126

Optimisation of 3D models for visualisation in virtual reality systems

Yurii Havryliuk*
PhD in Technical Sciences, Associate Professor
National Technical University "Kharkiv Polytechnic Institute"
61002, 2 Kyrpychova Str., Kharkiv, Ukraine
https://orcid.org/0000-0001-7762-5699

Abstract. The relevance of the study is conditioned by the growing use of virtual reality in education, medicine, design and engineering systems, where traditional 3D models are often too resource-intensive for stable interactive reproduction. The purpose of the study was to identify and test the effectiveness of complex adaptation of digital objects to interactive spatial reproduction in conditions of limited hardware resources. The research methodology combined theoretical analysis with experimental testing in Unity 2022.3 on the Meta Quest 2 platform with performance and visual quality assessment. The results of the study showed that direct use of traditional 3D models in virtual reality is ineffective due to excessive load on hardware resources and instability of scene reproduction. It was established that the highest efficiency is provided by complex adaptation, which combines simplification of the polygonal grid, reduction of the volume of texture data, simplification of materials, application of detail levels, and stage optimisation. After adaptation, the number of polygons in the studied models decreased by 95-96%, and the volume of texture data – by 85-88%. In the test virtual environment, the average frame rate increased from 42 to 85 frames per second, which confirmed a significant increase in scene performance. However, the increase in performance was not accompanied by a critical loss of visual quality, since the overall shape of objects remained at the level of 98-99%, and the level of visual immersion remained high. The results of the study can be used by developers of virtual reality applications, computer graphics specialists, teachers, designers, and engineers to create training simulators, architectural visualisation, medical simulators, and technical virtual reality systems with increased performance and visual realism

Keywords: application of detail levels; simplification of the polygonal grid; interactivity; visualisation; level of visual immersion; frame rate

INTRODUCTION

The growing use of digital immersive technologies in education, cultural heritage, design, advertising, and visual communications has highlighted the challenge of effectively utilising 3D models in environments where high frame rates, interactivity and visual realism are all required simultaneously. For such systems, the geometric detail of the object itself is no longer enough, since without special adaptation, even a high-quality model can create an excessive load on the graphic subsystem, worsen the stability of scene reproduction, and reduce user comfort. That is why the use of three-dimensional modelling in computer graphics is associated not only with the creation of visually complex content, but also with its rational adaptation to the requirements of interactive immersive environments. In the scientific Ukrainian discourse, the topic of three-dimensional content was actively developed primarily in the pedagogical context. S. Lytvynova (2022) showed that students' readiness to use virtual reality (VR) in the educational process depended not only on access to digital tools, but also on the development of the ability to work with immersive content as a special type of learning environment. In the subsequent research, S. Lytvynova (2024) found that 3D content and 360-degree formats expanded the

Article's History: Received: 10.11.2025; Revised: 09.03.2026; Accepted: 18.05.2026; Published: 26.06.2026

Suggested Citation:
Havryliuk, Yu. (2026). Optimisation of 3D models for visualisation in virtual reality systems. *Bulletin of Cherkasy State Technological University*, 31(2), 126-138. doi: 10.62660/bcstu/2.2026.126.

*Corresponding author (yuriihavryliuk011@outlook.com)

Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/)

teacher's didactic capabilities, but their effectiveness directly depended on the quality of digital materials preparation and their compliance with practical educational tasks. A.S. Derkach & I.A. Tverdokhlib (2024), analysing the state of study of three-dimensional modelling in the Ukrainian school, noted that this area has already gone beyond the narrow technical specialisation and acquired the value of basic digital competence related to engineering, medicine, manufacturing, and design. S.O. Semerikov *et al.* (2024) expanded on this context by proposing a training methodology for WebAR development incorporating machine learning models, which highlighted the increasing complexity of modern immersive educational environments. P.P. Nechypurenko *et al.* (2023) demonstrated the practical effectiveness of a virtual chemical laboratory based on augmented reality to support students' learning and research activities. Together, these studies confirmed the high relevance of three-dimensional and immersive environments in education, but did not provide a specific explanation of how detailed digital models should be adapted for stable operation in VR.

A similar trend was observed in papers devoted to digital spatial reconstructions, virtual expositions, and visual communications. I. Tytarenko *et al.* (2023) proved that digital instruments allowed reproducing historical environments with a high level of spatial accuracy, and three-dimensional modelling could act not only as a means of visualisation, but also as a tool for research reconstruction. N. Dumanskyi & N. Vovk (2025) stressed that virtual and augmented reality technologies have strengthened the possibilities of preserving, reproducing and popularising the cultural heritage of Ukraine, especially in the face of military threats. M. Chernikov *et al.* (2025), analysing multimedia virtual exhibition spaces, showed that not only technical characteristics were crucial for the digital environment, but also aesthetic integrity, navigation convenience, and emotional immersion of the user. M.S. Strelchuk (2025) showed that three-dimensional modelling in advertising products was used as one of the leading means of forming a visual image, and the combination of 3D graphics with VR, augmented reality (AR) technologies, animation, and gamification increased the emotional involvement of the audience. All these studies convincingly substantiated the value of three-dimensional content for education, culture, design, and communication, but the issue of its quantitatively verified optimisation for VR scenes remained insufficiently disclosed.

Another important area was formed at the intersection of three-dimensional scenes, digital interaction, and new ways of presenting spatial information. R. Partsey *et al.* (2025) found that text annotations expand the ability to interpret three-dimensional scenes and enhance their semantic clarity in complex digital environments. This indicated a growing interest not only in creating 3D objects, but also in their structural organisation and adaptation to new interaction formats. Together with above research papers, this pointed to a significant gap: despite a significant interest in the educational, cultural, design, and communication potential of three-dimensional content, Ukrainian studies did not pay enough attention to the complex adaptation of models specifically for interactive VR, where frame rate, stage stability, textural rationality, and the preservation of visual realism became critical.

Despite the scientific significance of the considered studies, they do not sufficiently cover the issues of complex adaptation of three-dimensional models to an interactive environment VR, in particular, a quantitatively verified combination of geometric, textural, material, and scene optimisation to ensure high performance, rendering stability and the preservation of visual realism simultaneously. The purpose of this study was to substantiate and experimentally test the effectiveness of complex adaptation of 3D models for interactive reproduction in a VR environment under conditions of limited hardware resources. The objectives of the study were: to systematise the main approaches to optimising 3D models for VR environments; to experimentally verify the impact of comprehensive model adaptation on performance, rendering stability, rendering latency and hardware load within a VR scene; to assess the preservation of the visual fidelity of the optimised models in terms of geometric similarity, texture rendering quality, perception of detail, and the level of visual immersion.

MATERIALS AND METHODS

The study was theoretical and empirical in nature and combined the analysis of scientific sources with practical experimental testing of 3D models in the VR environment. At the first stage of the study, the material of theoretical analysis involved the research papers by S. Sukaridhoto *et al.* (2023), A. Geris *et al.* (2024), R. Grande *et al.* (2024), X. Hu *et al.* (2025), G. Papaioannou *et al.* (2025), Y. Wang *et al.* (2025), and H. Zhang & C. Peng (2026), which considered contemporary approaches to optimising 3D content for the VR environment. The choice of these sources was determined by the fact that they covered the main areas of adaptation relevant for this study, namely, simplification of the polygonal grid, optimisation of textures and materials, use of detail levels, organisation of rendering, and stage optimisation. Based on these studies, the theoretical basis of experimental design was established. Within this stage, approaches to geometric optimisation, reducing texture load, simplifying materials and visual effects, applying levels of detail, and stage optimisation were summarised. The theoretical analysis became the basis for constructing an experimental design and selecting technical parameters for further testing the effectiveness of adapting 3D models.

The empirical part of the study was performed in the Unity 2022.3 environment with integration under the standalone VR platform Meta Quest 2. The original models were created in Blender 3.6 in a highly polygonal

format with maximum geometric detail and complex physically correct textures, and then imported into Unity without prior optimisation. Further adaptation was carried out using the built-in tools Mesh Optimiser, Texture Compression, Automatic Level of Detail (LOD) Generator, and Asset Store plugins focused on the requirements of autonomous VR systems. The research material consisted of four basic 3D models representing typical objects of virtual scenes: the chair model as a static interior element, the tree model as a landscape object, the building facade model as an architectural element, and the anthropomorphic character model as a dynamic object. This selection allowed covering various types of geometric complexity, texture saturation, and functional purpose of objects in a VR environment. At the first stage, all models were analysed in their original state. At the second stage, a complex adaptation was applied to them, which included simplifying the polygonal grid, creating 3-4 levels of detail, combining textures, using Adaptive Scalable Texture Compression (ASTC), replacing some highly detailed colour textures with normal maps and roughness maps, and simplifying materials and programmes for calculating visual effects.

To assess the impact of adaptation on performance, a test scene was created that contained 120 objects as a combination of four basic models and reproduced a loaded VR environment of a mixed "interior + landscape" type. The measurements were performed in standalone VR mode on Meta Quest 2 over 300-second sessions. Tests were conducted on a standalone Meta Quest 2 VR helmet in offline mode, without connecting to a personal computer to render the scene. The device's hardware platform included a Qualcomm Snapdragon XR2 system-on-chip, 6 GB of RAM, and a display with a resolution of 1832×1920 pixels per eye, and support for refresh rates of 60, 72, 80, 90 and 120 Hz. The experiment used a 72 Hz mode because it met the minimum target level of stable VR scene playback for a standalone device. The internal memory of the test device was 128 GB. During the experiment, third-party background applications were closed, and scenes were launched without streaming images from a personal computer. Graphics and CPU load, memory usage, frame rate, and latency were recorded using Unity Profiler and Oculus Performance Metrics. This configuration allowed evaluating the effectiveness of the models' adaptation specifically for a resource-constrained standalone VR platform. The data collection tools were Unity Profiler and Oculus Performance Metrics. Descriptive statistics were used to quantify the stability of scene reproduction, in particular, the average Frame Per Second (FPS) rate and the sample standard deviation of FPS values. The standard deviation (SD) was calculated based on a sequence of frame rate values recorded during each 300-second test session using Unity Profiler and Oculus Performance Metrics, using the equation (1):

$$SD = \sqrt{\frac{\sum_{i-1}^{n}(FPS_i - FPS)^2}{n-1}}, \quad (1)$$

where FPS_i – frame rate value in the i-th time frame; FPS – average frame rate value per session; n – number of recorded values. A lower standard deviation value was interpreted as higher rendering stability and lower variability in VR scene performance. The analysis covered the average frame rate, the minimum frame rate at the lowest values of 1%, graphics and CPU load, video memory usage, scene loading time, and the delay between user movement and image refresh. In addition, rendering stability, the proportion of frames with a constant frequency of at least 72 FPS, and the standard deviation of the frame rate were evaluated.

A combination of objective and subjective approaches was used to assess visual quality after adaptation. Objective analysis included preserving the overall shape of objects, silhouette characteristics, structural similarity of textures, perception of small details at different distances, integrity of the scene, and smooth transitions between levels of detail. For this purpose, the Chamfer Distance metric for geometric similarity and the Structural Similarity Index Measure (SSIM) metric for evaluating structural similarity of textures in Unity Editor were used. The subjective assessment was conducted in the format of a blind test with the participation of 10 people, including 5 developers of 3D and VR content and 5 users with at least 6 months of experience using VR applications. All participants previously had basic experience working with the Meta Quest 2 helmet, which reduced the impact of adaptation to the device on the evaluation results. Scene comparisons were performed individually under the same viewing conditions, and the order in which the original and optimised scenes were shown was randomised for each participant. Participants did not receive information about which version of the scene was shown at a particular moment, and after each viewing, they rated the level of visual immersion, preservation of the shape of objects, and overall perception of the environment on a five-point scale. This design of the study helped to simultaneously trace the impact of adapting 3D models on the technical performance of the VR system and on maintaining the realism of the scene.

RESULTS

In a VR environment, visualisation efficiency was associated with the need to simultaneously provide high frame rates, stereoscopic playback, and low power consumption, especially for mobile and standalone devices. In the current review of effective visualisation for VR, rendering with resource concentration in the focus area, stereoscopic rendering, cloud rendering, and energy-saving rendering were highlighted as key areas for reducing computing load, and the role of increasing resolution and intermediate frame generation technologies to improve the efficiency of mobile VR systems (Hu *et al.*, 2025). For

large three-dimensional polygonal models, adaptation to VR was mainly associated with a combination of levels of detail and an approach focused on the central viewing area. The study by H. Zhang & C. Peng (2026), dedicated to real-time visualisation of large three-dimensional models with a concentration of computational resources in the focus area, showed that when selecting the level of detail, priority can be given to high geometric accuracy in the central part of the field of view, while in other areas the complexity of the model decreased. This approach was proposed specifically for VR scenes with very large polygonal models, where conventional desktop visualisation techniques created significant computational difficulties after being ported to VR.

Maintaining visual realism in VR depends not only on the model's geometry, but also on the scene's lighting. The study by G. Papaioannou *et al*. (2025) on approximate dynamic global lighting for VR noted that lighting is an important component of visual immersion, whereas for typical autonomous VR systems or mobile platforms, dynamic lighting calculations remain computationally complex. The proposed approach was aimed specifically at mid- to entry-level graphics equipment, which confirmed the need to adapt the visual components of the scene to the limited resources of VR devices. At the level of stage organisation, the effectiveness of VR adaptation was associated with a set of technical solutions, and not with one separate technique. In the applied study by Y. Wang *et al*. (2025), the optimisation techniques included the reuse of a single graphical representation of objects, batching of rendering calls, simplification of the polygon mesh for levels of detail, and culling of objects based on the viewing direction; it was noted that it was precisely these measures that ensured a stable frame rate even in the presence of thousands of objects of the same type. This gave grounds to consider the effectiveness of 3D modelling in VR as a result of matching the geometry, structure of detail levels, and organisation of the scene visualisation process.

The difference between conventional 3D models and VR-adapted models was mainly evident in the performance requirements of the interactive environment. If conventional models focus primarily on high image quality, then VR models should provide a stable frame rate, low display latency, moderate load on the Graphics Processing Unit (GPU), and a rational texture and material structure of the scene. That is why even a visually high-quality model without optimisation is often not effective enough for use in VR. The key differences between conventional 3D models and VR-adapted models are shown in Table 1.

Table 1. Comparison of conventional and VR-adapted 3D models

Comparison criteria	Conventional 3D models	VR-adapted 3D models	Analytical assessment
Main purpose of creation	High quality of an individual frame, static or pre-calculated visualisation	Stable real-time scene playback	In VR, priority is shifted from maximum detail to interactivity and performance
Number of polygons	Often high, without strict restrictions	Reduced by simplifying the polygonal grid and optimising LOD	Reducing polygonality is one of the main ways to improve the model's suitability for VR
Textures	Large and numerous texture files can be used	Textures are optimised, combined, and some details are transferred to normal maps	Texture optimisation reduces memory and rendering load without critical loss of visual quality
Materials and shaders	Often complex, focused on maximum visual effect	Simplified, subject to real-time restrictions	In VR, it is not the maximum complexity of the material that is important, but its computational feasibility
Memory and GPU loads	Increased, especially in difficult scenes	Reduced by optimising digital objects	For standalone VR devices, this directly affects the performance of the scene
Frame rate (FPS)	May decrease when porting to VR	After optimisation, it increases; the gain may be moderate, but its value depends on the complexity of the scene, the type of device, and the depth of optimisation	VR adaptation is directly related to FPS stabilisation and enhancement
Use the GPU	Often higher due to the excessive complexity of digital objects	After optimisation, it decreases; the study recorded a decrease by 6.73% and 11.72%	This is especially important for autonomous VR systems with limited resources.
Visual quality	Saved as much as possible in an isolated view	Partially simplified, but retains an acceptable similarity	In a number of models, after optimisation, the reduction of polygons by more than 99% did not lead to a critical loss of perceived quality
Suitability for an interactive VR environment	Limited without additional adaptation	High with comprehensive optimisation	VR model is evaluated as part of the entire rendering system, not as an isolated object

Source: created by the author based on S. Sukaridhoto *et al.* (2023), A. Geris *et al.* (2024), R. Grande *et al*. (2024), X. Hu *et al*. (2025)

Analysis of the table showed that the key difference between conventional 3D models and VR-adapted models was the change in design priorities. In conventional computer graphics, the focus on maximum detail of shape, textures and materials prevails, while in VR, the ability of the model to function stably in an interactive environment without overloading the system has become crucial. These characteristics showed that VR adaptation included simultaneous reduction of polygonality, optimisation of textures, simplification of materials, and reduction of memory and GPU load. It was this comprehensive restructuring of the model that provided an increase in frame rate and rendering stability, which became critical for a comfortable perception of the virtual environment. However, the adaptation did not lead to a sharp deterioration in visual quality, since with the correct combination of geometric simplification with texture maps and normal maps, the model remained sufficiently realistic. It was particularly significant that the optimised models showed not only less resource load, but also better compliance with the technical limitations of autonomous VR devices. This confirmed that the suitability of the 3D model for VR was determined not by the absolute level of detail, but by the balance between visual realism and computational efficiency. Thus, the data confirmed that VR-adapted models were functionally more efficient than conventional ones, as they better met real-time requirements, provided higher scene stability, and allowed maintaining acceptable visualisation quality at lower hardware costs.

The main areas of optimisation of 3D models for VR covered several interrelated levels of adaptation, since the effectiveness of the virtual environment was determined not only by the geometric complexity of the object, but also by the texture load, lighting parameters, levels of detail, and the overall organisation of the scene. That is why the effective adaptation of models to VR was based on a complex combination of geometric, textural, visual, LOD, and scene optimisation, each of which performed a separate function in reducing computational costs and maintaining acceptable visual quality. It is advisable to present these relationships in the form of a diagram Figure 1.

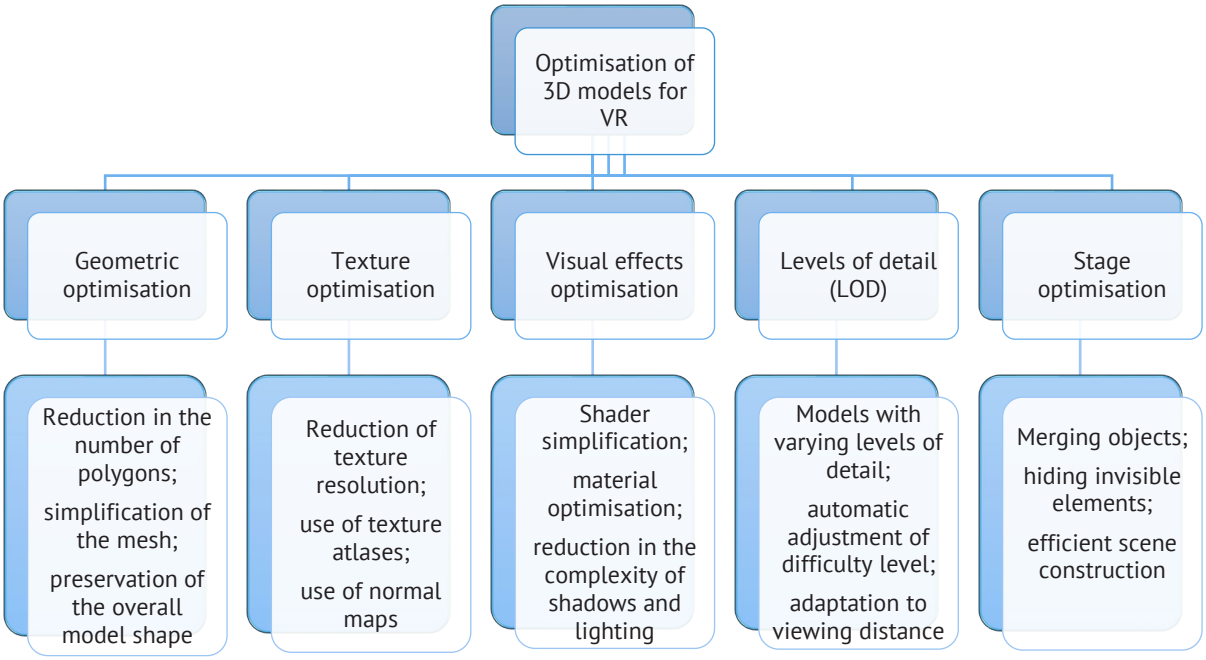


Figure 1. Main areas for optimising 3D models for VR

Source: created by the author based on R. Grande *et al.* (2024), Y. Wang *et al.* (2025), X. Hu *et al.* (2025), G. Papaioannou (2025), H. Zhang & C. Peng (2026)

Analysis of Figure 1. showed that the optimisation of 3D models for VR was complex and covered several interrelated areas, each of which performed a separate function in reducing the computational load. Geometric optimisation reduced the polygonality of the model, texture optimisation reduced the amount of graphic resources, and visual effects optimisation simplified shaders, materials, shadows, and lighting. The use of detail levels helped to dynamically change the complexity of the object depending on the conditions of perception, while scene optimisation covered the rational organisation of the entire virtual space. The data also confirmed that none of the above areas would have provided a sufficient result separately. The greatest effect was achieved precisely by combining several approaches consistent with the model type, scene parameters, and technical capabilities of the VR platform. Thus, the effectiveness of adapting 3D models to VR was determined not by a separate simplification method, but by a balanced application of a set of optimisation solutions.

The results of adapting basic 3D models to the VR environment were analysed sequentially in four key aspects: geometry, texture data, rendering, and overall scene stability. The most pronounced practical effect was observed in complex scenes where a significant number of highly detailed objects were used simultaneously. In the absence of optimisation, such a scene created a critical load on the GPU and video memory, which was accompanied by significant performance drops and reduced playback smoothness. After adaptation, even with a partial reduction in detail, the scene became suitable for full-fledged interactive interaction, while maintaining a sufficient level of visual realism. The most noticeable changes were recorded at the level of geometric optimisation. The use of simplification of the polygonal grid in combination with the development of 3-4 levels of detail (LOD) helped to significantly reduce the number of polygons without noticeably disrupting the overall shape of objects. This is especially important for VR environments where excessive polygonality directly affects rendering speed and frame rate stability. For a visual representation, the basic 3D models before and after optimisation are shown in Figure 2 and Figure 3.

Figure 2. Output high-poly 3D models prepared in Blender 3.6
Notes: model of a chair, a tree, a building facade, and an anthropomorphic character is presented.
Source: created by the author in Blender 3.6

Figure 3. View of basic 3D models after a comprehensive adaptation in Unity 2022.3 for the VR environment
Notes: overall silhouette and spatial proportions of models are preserved while simplifying geometry, textures, and materials
Source: created by the author

As can be seen from Figure 2, after adaptation, the models retained their basic shape, silhouette, and functional recognition, but became less detailed at the level of fine geometry, texture saturation, and material relief. The most noticeable simplification could be traced in the crown of the tree, decorative elements of the facade, and small details of the character, while the chair retained almost the same overall design. This type of change corresponded to the task of VR adaptation: not to replace the original objects with new models, but to

reduce their resource intensity while maintaining the visual realism of the scene. Quantitative changes in the main parameters of models before and after adaptation are shown in Table 2.

Table 1. Changes to the main parameters of basic 3D models before and after adapting to the VR environment

Model	Number of polygons (before)	Number of polygons (after)	Change, %	Texture size, MB (before)	Texture size, MB (after)	Change, %	Average object rendering time, ms (before)	Average object rendering time, ms (after)
Chair	248,750	12,438	-95	85	12	-86	1.2	0.4
Tree	1,456,320	58,253	-96	210	28	-87	3.5	1.1.
Building facade	3,245,800	162,290	-95	420	52	-88	8.2	2.9
Character	856,400	42,820	-95	150	22	-85	2.8	0.9

Source: created by the author based on testing in Unity3D

The data in Table 2 showed that in all the models studied, the number of polygons decreased by 95-96%, while the volume of texture data decreased by 85-88%. The greatest absolute reduction in polygonality was recorded for the building facade model, which was explained by the high initial geometric complexity of this object. Simultaneously, even after this simplification, the visual structure of the models remained sufficiently complete for perception in the VR scene. Texture data was optimised by combining textures, applying adaptive texture compression suitable for mobile VR platforms, and replacing some highly detailed colour textures with normal maps and roughness maps. This allowed reducing the total volume of textures by an average of 86.5%, which significantly lowered the load on video memory and contributed to more stable scene reproduction. At the visualisation level, a positive effect was achieved by simplifying materials and programmes for calculating visual effects. The transition from complex physically correct shading models to simplified lighting models with a minimum number of samples provided an additional reduction in the computational load. As a result, the average visualisation time of an individual object decreased in all cases: for a chair – from 1.2 to 0.4 ms, for a tree – from 3.5 to 1.1 ms, for a building facade – from 8.2 to 2.9 ms, for a character – from 2.8 to 0.9 ms. In relative terms, this corresponded to a reduction of approximately 64-67%, which directly affected the overall performance of the scene. The most significant changes were recorded at the level of overall stability of the VR environment. In the test scene, which contained 120 adapted objects, the average frame rate increased from an unstable 38-45 FPS, with drops to 25 FPS, to a stable 78-92 FPS. This showed that the adaptation turned resource-intensive conventional models into functionally efficient elements of the virtual environment, suitable for long-term interactive use without significant loss of realism.

To assess the impact of complex adaptation of 3D models on the functioning of the VR system, key performance indicators were compared before and after scene optimisation. During this stage, changes in frame rate, graphics and CPU load, video memory usage, scene loading time, and delays from motion to image update were analysed. This approach helped to quantify the extent to which geometric, textural, material, and stage optimisation tools were used to improve the suitability of the virtual environment for stable interactive use (Table 3).

Table 3. Comparison of key VR system performance indicators before and after comprehensive adaptation of 3D models

Indicator	Before adaptation	After adaptation	Change
Average FPS	42	85	+102%
Minimum FPS (1% low)	25	72	+188%
GPU load, %	94	47	-50%
CPU load, %	68	39	-43%
Device RAM/VRAM usage, MB	1,420	580	-59%
Scene loading time, s	8.4	2.9	-65%
Motion-to-photon latency, ms	38	19	-50%

Source: created by the author based on testing in Unity 2022.3 on the standalone Meta Quest 2 platform using Unity Profiler and Oculus Performance Metrics

Analysis of Table 3 showed that complex adaptation of 3D models had a positive effect not only on the average level of performance, but also on the stability of scene reproduction. Prior to optimisation, there were frequent frame rate drops below 30 FPS, covering up to 18% of the session duration and accompanied by marked intermittent movement and an increased risk of vestibular discomfort. After adaptation, the proportion of frames with a stable frequency of at least 72 FPS reached 96%, and the sample standard deviation of FPS values calculated from the data of a 300-second test session decreased from 12.4 to 3.8 (Equation 1).

This confirmed that for VR, it is important not only to increase the average FPS, but also to reduce performance fluctuations throughout the session. The greatest contribution to improving efficiency was provided by a combination of several optimisation tools. Generating detail levels combined with clipping objects that are insignificant for the current observer position reduced the GPU load by 32%. Combining textures and applying ASTC compression reduced video memory usage by 41%. Simplification of materials and shaders reduced the visualisation time of an individual object by an average of 67%. Batch processing of rendering calls and repeated use of a single graphical representation of the object had an additional positive effect in scenes with a large number of objects of the same type. It was found that excessive simplification of models reduces the quality of visual perception. In the test scenes, the reduction of polygonality by more than 98% in combination with the complete removal of normal maps led to a noticeable loss of small details already at a distance of 3-5 m: the expressiveness of the texture of facades, and the surface of tree bark decreased, which weakened the immersion effect. The most balanced mode was when the number of polygons was reduced by 94-96% while maintaining at least two levels of detail. Thus, the adaptation of 3D models turned out to be not an auxiliary, but a basic factor in ensuring the performance of VR systems, especially on equipment with limited resources. Comprehensive optimisation has ensured that the stable performance required for long-term comfortable use of the virtual environment is achieved, without critical loss of its visual credibility. It should be noted that FPS growth rates are variable and depend on the experimental conditions, scene structure, model complexity, amount of optimisation applied, hardware platform, and measurement method. Therefore, it is advisable to interpret the obtained values considering the specific configuration of the test environment.

Evaluation of visual realism after adapting 3D models to VR conditions showed that reducing geometric and textural complexity did not lead to a critical deterioration in scene perception. The analysis was performed on several aspects: preserving the overall shape of objects, the quality of texture reproduction, the perception of small details at different distances, the integrity of the scene, and the level of visual immersion. This approach helped to determine how much optimisation affected not only technical efficiency, but also the reliability of the virtual environment (Table 4).

Table 4. Evaluation of the visual realism of 3D models after optimisation for VR

Evaluation criteria	Result after optimisation	Analytical assessment
Preservation of the general shape of objects	98-99% by Chamfer Distance metric	Overall contours and spatial proportions of the models remained virtually unchanged
Stability of the object silhouette	Contours of the chair, wood, and facade of the building were preserved after the polygons were reduced by 95%	Geometric simplification did not violate the visual recognition of the main shapes
Texture playback quality	Average SSIM for textures – 0.94	Structural similarity of textures remained high
Transfer of surface microdetals	Partial loss of high-frequency details of diffuse textures was compensated by roughness maps and normals	Visual plausibility of the surfaces was preserved even when the texture volume was reduced
Perception of small details at close range	At a distance of up to 2 m, the differences were minimal	Optimised models remained convincing during typical object interactions
Perception of small details at medium and long distances	At a distance of 5-10 m, small elements were perceived as more smoothed	Loss of individual details did not destroy the overall impression of the scene
Smooth transitions between detail levels	Changes in the level of detail were almost not visually recorded	System of detail levels provided an imperceptible simplification of the model during the user's movement
Stage integrity	Preserved at a high level	Optimisation did not break the consistency of the visual environment
Level of visual immersion	Average rating – 4.7 out of 5	High stability of the stage increased the sense of presence in the virtual space

Source: created by the author based on testing in Unity

The data in Table 4 showed that the overall shape of objects and their silhouette characteristics were the most resistant to optimisation. The quality of texture reproduction remained high. Despite a partial decrease in the detail of colour textures, the structural similarity remained at 0.94, which indicated a slight loss of visual information. An additional role in maintaining the plausibility of surfaces was played by the use of normalcy maps and roughness maps, due to which it was possible to compensate for the simplification of texture resources without significantly impairing the overall perception of materials. It was found that the perception of small details significantly depended on the observation distance. At close range, which is typical

for direct interaction with objects, the differences between the basic and optimised models were minimal. At medium and long distances, individual elements, such as the texture of bark, leaves, or small architectural details, became more smoothed out, but this did not lead to a loss of the integrity of the visual image. In addition, this was facilitated by smooth transitions between levels of detail, which did not create noticeable gaps in perception. Separately, it should be noted that a high level of visual immersion was provided not only by the geometry or textures themselves, but also by the stability of scene reproduction. With an average score of 4.7 out of 5, the optimised environment was perceived as visually compelling and comfortable for long-term use. This showed that in VR conditions, realism is determined not by the maximum complexity of each individual object, but by the consistency of geometry, textures, lighting, and smooth reproduction. To specify the results of objective and subjective assessment of visual quality, numerical indicators of geometric similarity of models using the Chamfer Distance metric, structural similarity of textures using the SSIM metric, and the results of blind subjective testing on a five-point scale were summarised. Summary data for individual models and distribution of participants' average scores are shown in Table 5.

Table 5. Objective indicators of visual quality of optimised 3D models

Model	Shape retention based on Chamfer Distance, %	SSIM textures
Chair	99	0.95
Tree	98.2	0.93
Building facade	98.7	0.94
Character	98.5	0.94
Mean value	98.6	0.94

Source: created by the author based on testing in Unity Editor and the results of a blind survey of 10 participants

The data in Table 5 showed that complex adaptation of 3D models to the VR environment did not cause a critical deterioration in visual quality either in terms of objective or subjective indicators. According to the results of an objective assessment, the shape retention of models using the Chamfer Distance metric was in the range of 98.2-99%, and the average value was 98.6%, which confirmed the high geometric similarity of the optimised models to the original ones. The structural similarity of textures by the SSIM metric varied between 0.93 and 0.95 with an average value of 0.94, which showed a slight loss of texture information after optimisation. The highest values were recorded for the chair model, while slightly lower values were recorded for the tree model, which can be explained by a more complex irregular geometry and a large number of small peripheral elements. To summarise the results of subjective evaluation of the optimised VR scene on a five-point scale, the average values for individual criteria and the distribution of participants' ratings were determined, which are shown in Table 6.

Table 6. Results of subjective evaluation of the optimised VR scene on a 5-point scale, n = 10

Criteria	4 points, n (%)	5 points, n (%)	Average score
Preservation of the shape of objects	3 (30)	7 (70)	4.7
Level of visual immersion	4 (40)	6 (60)	4.6
General perception of the environment	2 (20)	8 (80)	4.8

Source: created by the author based on the results of blind testing

The results of subjective evaluation also confirmed the preservation of high visual realism of the optimised scene. According to all criteria, high ratings prevailed, and the share of maximum ratings ranged from 60% to 80%. Overall perception of the environment was rated the highest, while preserving the shape of objects and the level of visual immersion were slightly lower. This showed that the optimisation carried out ensured not only an increase in the technical efficiency of the VR scene, but also the preservation of its integrity, object recognition, and comfort for user perception. Thus, the data obtained confirmed that a rational combination of geometric simplification, texture optimisation, and the use of detail levels helped to achieve a balance between performance and visual plausibility of the virtual environment.

The results of adaptation of 3D models showed not only an increase in the technical efficiency of the virtual environment, but also the practical feasibility of using them in various types of VR applications. Reducing polygonality, reducing the amount of texture data, simplifying materials, and stabilising rendering helped to use complex scenes with limited hardware resources without significantly losing visual realism. One of the main advantages was the reduction of hardware requirements, which expanded the possibilities of using VR in educational, project and production environments. An important result was also an increase in the smoothness of interactive interaction and a reduction in the risk of visual discomfort. In addition, optimisation made it easier to scale complex scenes, as it allowed increasing the number of objects without proportionally

increasing the load on the system. Therefore, the practical value of optimised models was to increase the availability, stability, and functional suitability of VR solutions, which expanded the possibilities of their application implementation.

Generalisation of theoretical and practical results showed that direct use of conventional 3D models in the VR environment is ineffective, since models created without considering the requirements of interactive rendering form an excessive load on the graphic subsystem and reduce the stability of scene reproduction. The highest efficiency was achieved with a comprehensive adaptation that combined geometric optimisation, reduced texture data, simplified materials, applied detail levels, and stage optimisation. The most effective solution was not a separate technical solution, but a coordinated combination of several approaches in accordance with the type of object, its role in the scene, and the technical capabilities of the platform. It was found that the increase in performance was not accompanied by a critical loss of visual quality, and the balance between realism and speed remained a crucial factor in VR. This adaptation was particularly important for systems focused on mass or resource-limited equipment. Therefore, the use of 3D modelling in computer graphics for VR is most effective when models are initially designed or purposefully modified to meet the requirements of the interactive VR environment.

DISCUSSION

The results confirmed that the effectiveness of using 3D modelling in VR is determined not by the maximum geometric detail per se, but by the ability of the model to maintain visual credibility with a significantly reduced computational load. The study found that a comprehensive adaptation of models for the VR environment, which combined simplification of the polygonal grid, optimisation of textures, materials, detail levels and scenes, provided a sharp increase in performance without critical loss of perception quality. This result gives grounds to consider the optimisation of 3D content not as an auxiliary technical stage, but as a key condition for the functional suitability of digital objects for an interactive VR environment.

Performance improvements were achieved through local simplification of geometry, textures, materials, and stage structure for the limited resources of an autonomous device. J.A. Santos-Paz *et al.* (2025) solved a similar problem through remote visualisation, data transfer to the server, and a low-latency codec that is resistant to packet loss and is associated with data about the user's gaze direction. Thus, both approaches were aimed at ensuring smooth VR interaction, but in the first case, the key tool was content optimisation, and in the second – network computing architecture. After optimisation, the overall shape of objects remained at the level of 98-99%, which indicated that there was no critical loss of visual integrity. A similar pattern was found by Z. Yang *et al.* (2025), which, when optimising BIM- and 3D GIS models for digital doubles, reduced the average load time by 15% while maintaining 97% geometric accuracy. The difference was that their study focused on the accuracy of the spatial representation of the digital twin, while here the priority remained to provide real-time in VR. The data obtained confirmed that even within the conventional polygonal approach, a significant increase in technical efficiency was achieved due to a rational combination of geometric, texture, and stage optimisation. This correlated with the review by W. Chen *et al.* (2025), which considered modern techniques of 3D reconstruction and visualisation, in particular the Gaussian splatting-reproduction approach, as a direction for improving the speed and quality of reproduction of complex scenes. Simultaneously, the review paper outlined the general evolution of methods, while the presented results showed the practical effectiveness of classical model optimisation in an applied VR environment. Increasing the frame rate and reducing latency was accompanied by maintaining a high level of visual immersion, i.e., technical optimisation directly improved the user's perception of the scene. The study by S. Fowler & S. Souihi (2025) also emphasised that the quality of the VR experience was determined by the balance between latency and visual clarity, and its optimisation was achieved through reinforcement learning on mobile edge networks. Thus, both studies confirmed the multi-factorial nature of VR performance, but in one case the decisive factor was the preparation of 3D content, and in the other – the adaptive distribution of network and computing resources.

The results showed that optimisation of three-dimensional models was an important prerequisite for their effective use in applied VR systems. This was consistent with research by Z. Li *et al.* (2025), where VR was used as an interactive design environment. The difference was that there the emphasis was placed on the development of engineering solutions, while here – on the technical suitability of models for stable scene reproduction. A similar applied orientation was traced in the study by A. Hernández-Chávez *et al.* (2026), dedicated to visualising the network infrastructure of university campuses in mobile applications and VR. Their results confirmed that 3D modelling in infrastructure systems performed not only a demonstration, but also an analytical function. Simultaneously, this study was dominated by the visualisation component, rather than productivity optimisation. The importance of the technical quality of the VR environment for practical application was also confirmed in the study by K. Wang *et al.* (2025), where VR was seen as a safety management tool in construction. The researchers attributed its effectiveness to the ability to model professional situations and improve risk recognition. In both cases, VR acted as an effective engineering tool, although the functional purpose of its application was different. A close example of the technical use of VR was provided by L.S. de Assis *et al.* (2026), who applied photogrammetric three-dimensional models and VR to

inspect tunnels. This confirmed that the combination of 3D modelling with VR was promising not only for design, but also for monitoring the state of infrastructure facilities. Thus, the results obtained fit well into the broader field of applied use of VR in engineering and technical systems.

The results showed that the value of a virtual environment was determined not only by technical stability, but also by the ability to maintain a holistic perception of the scene and user-friendly interaction. This was consistent with the research by F. Savini *et al*. (2025), in which 3D models based on unmanned photogrammetry and laser scanning were semantically enriched, simplified, and integrated into a web-based VR solution to increase the accessibility of cultural heritage. Thus, in both cases, optimisation of 3D content was considered as a condition not only for technical suitability, but also for greater accessibility of cultural experience. The educational potential of the immersive environment was confirmed in the study by S.-J. Chen *et al*. (2024), where an immersive VR approach to 3D modelling increased creativity and problem-solving abilities in younger students and reduced cognitive load compared to the usual virtual format. A similar effect was found by I. Irwanto *et al*. (2025), who recorded a statistically significant improvement in the creative attitudes and academic performance of 11th grade pupils. This showed that a well-organised immersive environment increased not only engagement, but also cognitive performance.

The importance of user interaction was clearly traced in the study by W. Zhao *et al*. (2023), where VR-based 3D modelling technology and a human-machine interaction algorithm for digital control and presentation of exhibits were proposed for the museum space. A logical approach was presented by E. Kontogiorgakis *et al*. (2024), who combined 3D reconstruction, virtual characters, and 360° video in a gamified VR narrative for cultural tourism. Thus, in cultural and museum applications, not only visibility became important, but also the ability of VR to form engagement and emotionally rich interaction. A broader theoretical context of this area was presented by A. Marougkas *et al*. (2023), who in a review of educational VR studies showed that the effectiveness of such environments was most often explained by active learning, experience-based learning, and increased engagement. In this sense, the results presented fit well with the general trend that VR was considered not just as a means of visualisation, but as a medium of perception, interaction, and cognition. Therefore, this block of research confirmed that in culture, education, and user experience, immersion was an independent factor in the effectiveness of the digital space. Thus, contemporary research has confirmed that the effectiveness of VR environments was determined not only by the technical quality of 3D models, but also by their ability to provide cognitive engagement, user-friendly interaction, and a holistic cultural user experience.

CONCLUSIONS

As a result of the study, it was found that direct use of conventional 3D models in a VR environment is ineffective, since models created without considering the requirements of interactive rendering create an excessive load on graphics and central processors and video memory. It was confirmed that the highest efficiency is achieved with a comprehensive adaptation that combines geometric optimisation, reducing the volume of texture data, simplifying materials, applying detail levels and stage optimisation. Practical testing showed that after adaptation, the number of polygons in the studied models decreased by 95-96%, and the volume of texture data – by 85-88%. The average visualisation time of an individual object was reduced by about 64-67%, which provided a significant increase in scene performance. In the VR test environment, the average frame rate increased from 42 to 85 frames per second, the minimum frame rate at 1% of the lowest values increased from 25 to 72 FPS, the GPU load decreased by 50%, video memory usage – by 59%, and the delay between user movement and image refresh – from 38 to 19 ms. This confirmed that the adaptation of 3D models is one of the key factors for ensuring the stable functioning of VR systems, especially on autonomous and resource-limited devices.

The visual quality assessment showed that the performance improvement was not accompanied by a critical loss of realism. The overall shape of the objects remained at 98-99%, the structural similarity of textures was 0.94, and the level of visual immersion was estimated at an average of 4.7 out of 5. This suggests that for VR, it is not the maximum detail of each object that is crucial, but the balance between geometric complexity, texture saturation, lighting, and playback stability. The results showed that even with a significant reduction in polygonality and the volume of texture data, the virtual environment retained the integrity of visual perception and remained convincing to the user. This showed that rational optimisation did not reduce the quality of the VR scene to an unacceptable level, but, on the contrary, provided a combination of technical efficiency with a fairly high visual plausibility. A limitation of the study was the use of a single hardware platform, a limited set of basic models, and a single test scene configuration. Further research should be directed to testing the results on other VR devices, expanding the types of virtual scenes, and applying adaptive real-time optimisation algorithms.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Chen, S.-J., Chen, C.-Q., & Shan, X.-F. (2024). The effects of an immersive virtual-reality-based 3D modeling approach on the creativity and problem-solving tendency of elementary school students. *Sustainability*, 16(10), article number 4092. doi: 10.3390/su16104092.

[2] Chen, W., Li, Z., Guo, J., Zheng, C., & Tian, S. (2025). Trends and techniques in 3D reconstruction and rendering: A survey with emphasis on Gaussian splatting. *Sensors*, 25(12), article number 3626. doi: 10.3390/s25123626.

[3] Chernikov, M., Vasylieva, O., Batrak, V., & Shaura, A. (2025). Aesthetic issues in the design of multimedia virtual exhibition spaces. *Art and Design*, 8(3), 22-35. doi: 10.30857/2617-0272.2025.3.2.

[4] de Assis, L.S., Poncetti, B.L., Machado, L.B., & Futai, M.M. (2026). Drone-based photogrammetry and virtual reality: Technological alternatives for tunnel inspection. *Tunnelling and Underground Space Technology*, 168(1), article number 107094. doi: 10.1016/j.tust.2025.107094.

[5] Derkach, A.S., & Tverdokhlib, I.A. (2024). The research of the state of studying 3D modeling in general secondary education schools of Ukraine. *Problems of the Modern Textbook*, 33, 106-116. doi: 10.32405/2411-1309-2024-33-106-116.

[6] Dumanskyi, N., & Vovk, N. (2025). Implementation of VR and AR technologies in the preservation of Ukraine's cultural heritage. *Library Science. Record Studies. Informology*, 4, 144-151. doi: 10.32461/2409-9805.4.2025.351584.

[7] Fowler, S., & Souihi, S. (2025). Multi-user QoE optimization in virtual reality gaming using RL on mobile edge networks. *Cluster Computing*, 28, article number 597. doi: 10.1007/s10586-025-05271-7.

[8] Geris, A., Cukurbasi, B., Kilinc, M., & Teke, O. (2024). Balancing performance and comfort in virtual reality: A study of FPS, latency, and batch values. *Software: Practice and Experience*, 54(12), 2336-2348. doi: 10.1002/spe.3356.

[9] Grande, R., Albusac, J., Vallejo, D., Glez-Morcillo, C., & Castro-Schez, J.J. (2024). Performance evaluation and optimization of 3D models from low-cost 3D scanning technologies for virtual reality and metaverse e-commerce. *Applied Sciences*, 14(14), article number 6037. doi: 10.3390/app14146037.

[10] Hernández-Chávez, A., López-Yáñez, I., Hernández-Chávez, M., & Fabila-Bustos, D.A. (2026). Development of mobile applications and virtual reality with 3D modeling for the visualization of network infrastructures on university campuses. *Technologies*, 14(3), article number 149. doi: 10.3390/technologies14030149.

[11] Hu, X., Wu, X., Ma, M., Xu, X., Gu, Y., Wang, G., Xu, Y., Meng, X., & Wang, L. (2025). Efficient VR rendering: Survey on foveated, stereo, cloud, and low-power rendering techniques. *Virtual Reality & Intelligent Hardware*, 7(5), 421-452. doi: 10.1016/j.vrih.2025.08.003.

[12] Irwanto, I., Lintangnicita, T., Cahyana, U., & Cha, J. (2025). Examining the effect of immersive virtual reality on 11th-grade students' creative thinking disposition and academic achievement. *Discover Education*, 4, article number 438. doi: 10.1007/s44217-025-00880-1.

[13] Kontogiorgakis, E., Zidianakis, E., Kontaki, E., Partarakis, N., Manoli, C., Ntoa, S., & Stephanidis, C. (2024). Gamified VR storytelling for cultural tourism using 3D reconstructions, virtual humans, and 360° videos. *Technologies*, 12(6), article number 73. doi: 10.3390/technologies12060073.

[14] Li, Z., Lee, T.-U., & Xie, Y.M. (2025). Interactive 3D structural design in virtual reality using preference-based topology optimization. *Computer-Aided Design*, 180, article number 103826. doi: 10.1016/j.cad.2024.103826.

[15] Lytvynova, S. (2022). Ready of students of general secondary education institutions to use virtual reality in the educational process. *Prospects and Innovations of Science*, 4(9), 218-230. doi: 10.52058/2786-4952-2022-4(9)-218-230.

[16] Lytvynova, S. (2024). 3D content in the educational practice of a general secondary education teacher. *Scientific Bulletin of Uzhhorod University. Series: "Pedagogy. Social Work"*, 1(54), 97-105. doi: 10.24144/2524-0609.2024.54.97-105.

[17] Marougkas, A., Troussas, C., Krouska, A., & Sgouropoulou, C. (2023). Virtual reality in education: A review of learning theories, approaches and methodologies for the last decade. *Electronics*, 12(13), article number 2832. doi: 10.3390/electronics12132832.

[18] Nechypurenko, P.P., Semerikov, S.O., & Pokhliestova, O.Y. (2023). An augmented reality-based virtual chemistry laboratory to support educational and research activities of 11th grade students. *Educational Dimension*, 8, 240-264. doi: 10.31812/educdim.4446.

[19] Papaioannou, G. (2025). Approximate dynamic global illumination for VR. *Virtual Reality*, 29, article number 54. doi: 10.1007/s10055-025-01114-3.

[20] Partsey, R., Teslyuk, V., Maksymets, O., Humennyy, V., & Kuzma, V. (2025). Enhancing 3D scene understanding via text annotations. *Eastern-European Journal of Enterprise Technologies*, 1(133), 76-84. doi: 10.15587/1729-4061.2025.323757.

[21] Santos-Paz, J.A., Garcia Aranda, J.J., Alarcón Granero, M., Quintanilla, F.J., Perez-Tirador, P., Otero, A., & Garcia-Carmona, R. (2025). NajVR: Low-latency plug-and-play remote rendering for interactive virtual reality applications. *Computing*, 107(12), article number 232. doi: 10.1007/s00607-025-01589-7.

[22] Savini, F., Cordisco, A., Fabbrocino, G., Giallonardo, M., Trizio, I., & Marra, A. (2025). From drone-based 3D model to a web-based VR solution supporting cultural heritage accessibility. *Drones*, 9(11), article number 775. doi: 10.3390/drones9110775.
[23] Semerikov, S.O., Foki, M.V., Shepiliev, D.S., Mintii, M.M., Mintii, I.S., & Kuzminska, O.H. (2024). Teaching WebAR development with integrated machine learning: A methodology for immersive and intelligent educational experiences. *Educational Dimension*, 10, 198-234. doi: 10.55056/ed.660.
[24] Strelchuk, M.S. (2025). Specificities of using 3D modeling in the process of creating visual materials in advertising products. *Ukrainian Art Discourse*, 6, 215-222. doi: 10.32782/uad.2025.6.25.
[25] Sukaridhoto, S., Haz, A.L., Fajrianti, E.D., & Budiarti, R.P. (2023). Comparative study of 3D assets optimization of virtual reality application on VR standalone device. *International Journal on Advanced Science, Engineering and Information Technology*, 13(3), 999-1008. doi: 10.18517/ijaseit.13.3.18375.
[26] Tytarenko, I., Pavlenko, I., & Dreval, I. (2023). 3D modeling of a virtual built environment using digital tools: Kilburun Fortress case study. *Applied Sciences*, 13(3), article number 1577. doi: 10.3390/app13031577.
[27] Wang, K., Antwi-Afari, M.F., Ossei-Gudom, B.M., Yi, W., Li, J., & Banihashemi, S. (2025). Recent advances in virtual reality applications for construction safety management. *Construction Innovation: Information, Process, Management*. doi: 10.1108/CI-11-2024-0357.
[28] Wang, Y., Zhang, E., Yang, A., Du, K., & Gao, J. (2025). Mixed reality-based multi-scenario visualization and control in automated terminals: A middleware and digital twin driven approach. *Buildings*, 15(21), article number 3879. doi: 10.3390/buildings15213879.
[29] Yang, Z., Aihemaiti, M., Abudureheman, B., & Tao, H. (2025). High-precision optimization of BIM-3D GIS Models for digital twins: A case study of Santun river basin. *Sensors*, 25(15), article number 4630. doi: 10.3390/s25154630.
[30] Zhang, H., & Peng, C. (2026). Performance-driven foveated VR rendering system for large 3D meshes. *Virtual Reality*, 30, article number 82. doi: 10.1007/s10055-026-01316-3.
[31] Zhao, W., Su, L., & Dou, F. (2023). Designing virtual reality based 3D modeling and interaction technologies for museums. *Heliyon*, 9(6), article number e16486. doi: 10.1016/j.heliyon.2023.e16486.

Оптимізація 3D-моделей для візуалізації в системах віртуальної реальності

Юрій Гаврилюк

Кандидат технічних наук, доцент
Національний технічний університет «Харківський політехнічний інститут»
61002, вул. Кирпичова, 2, м. Харків, Україна
https://orcid.org/0000-0001-7762-5699

Анотація. Актуальність дослідження зумовлена зростанням використання віртуальної реальності в освіті, медицині, проєктуванні та інженерних системах, де традиційні 3D-моделі часто виявляються надто ресурсоємними для стабільного інтерактивного відтворення. Метою дослідження було виявлення й перевірка ефективності комплексної адаптації цифрових об'єктів до інтерактивного просторового відтворення за умов обмежених апаратних ресурсів. Методологія дослідження поєднувала теоретичний аналіз із експериментальним тестуванням у Unity 2022.3 на платформі Meta Quest 2 з оцінюванням продуктивності та візуальної якості. Результати дослідження показали, що пряме використання традиційних тривимірних моделей у віртуальній реальності є малоефективним через надмірне навантаження на апаратні ресурси та нестабільність відтворення сцени. Встановлено, що найвищу ефективність забезпечує комплексна адаптація, яка поєднує спрощення полігональної сітки, скорочення обсягу текстурних даних, спрощення матеріалів, застосування рівнів деталізації та сценову оптимізацію. Після адаптації кількість полігонів у досліджених моделях зменшилася на 95-96 %, а обсяг текстурних даних – на 85-88 %. У тестовому віртуальному середовищі середня частота кадрів зросла з 42 до 85 кадрів за секунду, що підтвердило істотне підвищення продуктивності сцени. Водночас підвищення швидкодії не супроводжувалося критичною втратою візуальної якості, оскільки загальна форма об'єктів збереглася на рівні 98-99 %, а рівень візуального занурення залишився високим. Результати дослідження можуть бути використані розробниками додатків віртуальної реальності, фахівцями з комп'ютерної графіки, викладачами, проєктувальниками та інженерами для створення навчальних симуляторів, архітектурної візуалізації, медичних тренажерів і технічних систем віртуальної реальності із підвищеною продуктивністю та збереженням візуальної переконливості

Ключові слова: застосування рівнів деталізації; спрощення полігональної сітки; інтерактивність; візуалізація; рівень візуального занурення; частота кадрів

ВІСНИК
Черкаського державного технологічного університету

Науковий збірник

Том 31, № 2. 2026

Заснований у 1996 р. Виходить чотири рази на рік

Оригінал-макет видання виготовлено у редакційно-видавничому відділі
Черкаського державного технологічного університету

Відповідальний редактор:
А. Лавданський

Підписано до друку 18 травня 2026 р.
Формат 60*84/8
Умов. друк. арк. 16,3
Наклад 50 прим.

Адреса видавництва:
Черкаський державний технологічний університет
18006, бульв. Шевченка, 460, м. Черкаси, Україна
Тел.: +380638013283
E-mail: info@bulletin-chstu.com.ua
https://bulletin-chstu.com.ua/uk

BULLETIN
of Cherkasy State Technological University

Scientific Collection

Volume 31, No. 2. 2026

Founded in 1996. Published four times per year

The original layout of the publication is made in the editorial and publishing department
of Cherkasy State Technological University

Managing editor:
A. Lavdanskyi

Signed for print of Mat 18, 2026.
Format 60*84/8
Conventional printed pages 16.3
Circulation 50 copies

Editors Office Address:

Cherkasy State Technological University
18006, 460 Shevchenko Blvd., Cherkasy, Ukraine
Tel.: +380638013283
E-mail: info@bulletin-chstu.com.ua
https://bulletin-chstu.com.ua/en