МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ІНЖЕНЕРІЯ

Науково-технічний журнал

Том 22, № 3

2025

ВІННИЦЯ
2025

ISSN 1999-9941
e-ISSN 2078-6387

Засновник:
Вінницький національний технічний університет

Рік заснування:
2004

*Рекомендовано до друку та поширення
через мережу Інтернет Вченою Радою
Вінницького національного технічного університету
(протокол № 8 від 23 грудня 2025 р.)*

Державна реєстрація: Ідентифікатор медіа R30-01507.
Рішення Національної Ради України з питань телебачення і радіомовлення
№ 1234, протокол № 25 (31.10.2023 р.).

Журнал входить до переліку наукових фахових видань України
Категорія: Б. Науки: технічні. Спеціальності: F2 (121) – Інженерія програмного забезпечення; F3 (122) – Комп'ютерні науки; F7 (123) – Комп'ютерна інженерія; F4 (124) – Системний аналіз та наука про дані; F5 (125) – Кібербезпека та захист інформації; F6 (126) – Інформаційні системи та технології; G6 (152) – Метрологія та інформаційно-вимірювальна техніка; G22 (163) – Біомедична інженерія
(наказ МОН № 409 від 17.03.2020 року).

Журнал представлено у міжнародних наукометричних базах даних, репозитаріях та пошукових системах:
НБУ ім. В. І. Вернадського, Polska Bibliografia Naukowa, OUCI (Open Ukrainian Citation Index), DOAJ, J-Gate, Ulrichsweb Global Serials Directory, Litmaps

Адреса редакції:
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
Тел: +38 (0432) 560848
Факс: +38 (0432) 465772
E-mail: info@itce.vn.ua
https://itce.vn.ua/uk

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
VINNYTSIA NATIONAL TECHNICAL UNIVERSITY

INFORMATION TECHNOLOGIES AND COMPUTER ENGINEERING

Scientific and Technical Journal

Vol. 22, No. 3
2025

VINNYTSIA
2025

ISSN 1999-9941
e-ISSN 2078-6387

Founder:
Vinnytsia National Technical University

Year of foundation:
2004

Recommended for printing and distribution
via the Internet by Vinnytsia National Technical University
(Minutes No. 8 of December 23, 2025)

State Registration:
Media identifier R30-01507
Decision of the National Council of Television
and Radio Broadcasting of Ukraine
No. 1234, Minutes No. 25, dated 31.10.2023.

The journal is included in the List of Scientific Professional Publications of Ukraine
Category "B". Spetialities: 0588 – Inter-disciplinary programmes and qualifications involving natural sciences, mathematics and statistics; 0612 - Database and network design and administration; 0613 - Software and applications development and analysis; 0688 – Inter-disciplinary programmes and qualifications involving Information and Communication Technologies; 0714 - Electronics and automation; 0788 – Inter-disciplinary programmes and qualifications involving engineering, manufacturing and construction
(Order of the Ministry of Education and Science No. 409 of 17.03.2020).

The journal is presented international scientometric databases, repositories and scientific systems:
Vernadsky National Library of Ukraine, Polska Bibliografia Naukowa,
OUCI (Open Ukrainian Citation Index), DOAJ, J-Gate, Ulrichsweb Global Serials Directory,
Litmaps

Editor's office address:
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
Telephone: +38 (0432) 560848
Fax: +38 (0432) 465772
E-mail: info@itce.vn.ua
https://itce.vn.ua/en

Редакційна колегія

Editorial Board

ЗМІСТ

CONTENTS

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 02.09.2025 Revised: 10.11.2025 Accepted: 23.12.2025

UDC 004.056.5:004.738.5:004.942 DOI: 10.31649/vitce/3.2025.09

Tamper-resistant architecture of Server-Driven UI with real-time Merkle proof verification

Vladyslav Ananchenko*

Postgraduate Student
Academician Stepan Demianchuk International University of Economics and Humanities
33000, 4 Stepan Demyanchuk Str., Rivne, Ukraine
https://orcid.org/0009-0004-8963-775X

Yurii Lotiuk

PhD in Pedagogical Sciences, Associate Professor
Academician Stepan Demianchuk International University of Economics and Humanities
33000, 4 Stepan Demyanchuk Str., Rivne, Ukraine
https://orcid.org/0000-0001-6696-5583

Abstract. Server-driven user interface systems require protection from unauthorised modifications to ensure the integrity and security of displayed data. The purpose of this study was to develop a cryptographically verifiable change log of the user interface for systems with a Server-Driven User Interface. Within the study, methods of theoretical modelling, experimental testing, software implementation, and analysis of the regulatory framework were applied to design, verify, and evaluate a cryptographic change log in a client-interface environment. The main results showed that the use of signed structured interface blocks with hashing and digital signatures ensured the impossibility of undetected interface modification on the client side. Construction of the change log based on a hash tree guaranteed authenticity, immutability, and cryptographic verification of each interface element even under complex distributed conditions. Integration with advanced React rendering mechanisms enabled real-time verification of interface authenticity, ensuring compliance with international standards for personal data protection and transaction security. Furthermore, the results showed that client verification of Merkle proofs for blocks in React detected modifications before rendering, with an average verification time of 0.328 milliseconds per block. Auditing of blueprint file changes and the publish-subscribe system ensured data traceability and relevance, while component rendering after updates lasted only 2.7 milliseconds for the main component and 0.4 milliseconds for the button. Experiments confirmed a 94% attack-blocking rate, a reduction in rendering latency (from 850 to 300 milliseconds under slow network conditions), and a cache hit rate maintained at 94-95% under low load. Combined with improved key interface-interaction metrics, these results demonstrate the effectiveness of the proposed architecture. The obtained findings may be used by developers of critical web applications to implement secure interfaces that verify integrity in real time and comply with international security requirements

Keywords: cryptographically verifiable change log; principle of non-repudiation; evolution of React; overhead evaluation; cache hit rate

Introduction

The Server-Driven User Interface (SDUI) assumes dynamic generation of interface structures based on configuration data received from the server. This approach enables centralised adaptation of the interface to the user context (role, language, access rights, etc.), but simultaneously introduces new risks: the client visualises data whose authenticity it cannot independently verify. Vulnerability to attacks through configuration tampering, supply-chain compromise, or internal server errors threatens the integrity and security of critical interfaces. The research problem lies in the absence of mechanisms that allow the client side to independently verify the authenticity of received interface

Suggested Citation:
Ananchenko, V., & Lotiuk, Yu. (2025). Tamper-resistant architecture of Server-Driven UI with real-time Merkle proof verification. *Information Technologies and Computer Engineering*, 22(3), 9-22. doi: 10.31649/vitce/3.2025.09

*Corresponding author

structures before rendering, without relying on delivery infrastructure or server environment. This creates a protection gap in dynamic interfaces, which must be closed by integrating verifiable integrity verification mechanisms.

The study by A.J. Undirwadkar (2025) emphasised the advantages of the SDUI architecture, particularly its ability to provide dynamic interface updates without redeployment, accelerating development cycles and ensuring cross-platform consistency. A. Orynchak *et al*. (2024) explored JavaScript capabilities for real-time monitoring and threat response, demonstrating the efficiency of logging in protecting web applications – relevant for client-side interface integrity verification. Moreover, L. Shport (2025) analysed payment-system vulnerabilities and proposed a cryptographic architecture with encryption, authentication, and auditing modules ensuring multi-level protection of dynamic interfaces.

H.B. Azhar *et al*. (2025) proposed a post-quantum structure – Quantum-Resistant Merkle Trees – combining Zero-Knowledge Scalable Transparent Argument of Knowledge, lattice cryptography, and randomised hash functions to enhance security and performance. Experiments demonstrated a 28-32% reduction in proof-generation time compared to classical Merkle trees while maintaining logarithmic verification complexity. The results of O. Patel (2022) showed that Merkle proofs ensured transaction integrity with zero disclosure, reducing computational costs and increasing confidentiality in distributed systems. In turn, P. Du *et al*. (2022) proposed a tamper-resistant data-query model based on B+ and Merkle trees, ensuring the authenticity of query results in blockchain systems and improving reliability and security of dynamic data. Furthermore, S. Ridhorkar & S. Mishra (2024) developed a secure blockchain-based system for digital asset and inheritance management using Quantum-Resistant Dilithium Signatures and Merkle trees to ensure immutability, transparency, and protection from quantum-computer attacks, strengthening trust in automated processes within Decentralised Applications.

In addition, M. Havatiuk & I. Saiapina (2025) proposed an efficient graphical-interface update mechanism based on reactive programming, virtual Document Object Model (DOM), and centralised state management, which reduced computational costs and improved update speed. In the study by O. Kuznetsov *et al*. (2025), a secure generation and verification system for QR codes was developed using digital watermarks and a neural-network authentication model, ensuring high accuracy in forgery detection in dynamic environments. Moreover, M.T. Rubel *et al*. (2025) examined the use of blockchain technologies for automating real-time financial reconciliation, reducing error risks and meeting regulatory requirements such as those of the Securities and Exchange Commission and the Public Company Accounting Oversight Board.

Thus, existing studies do not address mechanisms for client-side interface-integrity verification before rendering in SDUI using signed JSON (JavaScript Object Notation) and Merkle proofs combined with React and change auditing. Therefore, the present study aimed to create a Cryptographically Verifiable Change Log (CVCLog) for user interfaces in SDUI architecture systems. The research objectives included proposing a modification-resistant interface architecture based on signed JSON packets organised into a Merkle tree, as well as implementing client-side integrity-proof verification before rendering in React and API (Application Programming Interface) level interface-change auditing.

Materials and Methods

This study, conducted in June-July 2025, used methods of theoretical modelling, experimental research, software implementation, and systematic analysis of the regulatory and legal framework. Using the JavaScript language and the Online JavaScript Compiler (Editor) – Programiz platform, a basic scheme was implemented for generating a cryptographically signed user-interface element in JSON format. To ensure controlled integrity, a Secure Hash Algorithm (SHA-256) hash function was calculated, and to confirm authenticity, the hash was signed using the Hash-based Message Authentication Code (HMAC) algorithm and a symmetric secret key. The same platform and language were also used to demonstrate an example of a Redis Pub/Sub message about cache invalidation. Theoretical foundations for ensuring UI-content authenticity in CVCLog were analysed by components – Merkle tree, principle of non-repudiation, digital signature, Access Control List (ACL), and Conflict-free Replicated Data Type (CRDT) (Badra & Borghol, 2018; Cai *et al*., 2022; Almeida, 2024). It was emphasised that SDUI directly transmitted ready-made React components through React server components.

Using the R language and the RStudio environment, a diagram of React rendering evolution (from React 16 to React 19) was constructed. The diagram visualised the chronological development of React, with major releases and key technological innovations of each version. The regulatory compliance of the CVCLog architecture was analysed for the Payment Card Industry Data Security Standard (PCI-DSS) (Agarwal *et al*., 2024), Payment Services Directive 2 (PSD2) (Christensen, 2025), and General Data Protection Regulation (GDPR) (Sienkiewicz, 2025). In addition, examples of security incidents were provided, such as the 7-Zip vulnerability in Ukraine (Girnus, 2025) and the data breach in the Episource medical company in the USA (Fadilpašić, 2025). These examples illustrated the potential consequences of insufficient data and interface protection, highlighting the necessity of cryptographic verification and auditing in the CVCLog architecture to ensure compliance with security standards such as PCI-DSS, PSD2, and GDPR.

Using the StackBlitz platform, the React library, and JavaScript, client-side Merkle proof verification for a JSON block in React was implemented. The crypto-js library was

used for cryptographic computations, performing SHA-256 hashing required for Merkle-proof generation and validation. The verification logic was implemented as a function that sequentially combined the hashes of the input JSON block with proof elements, forming a final hash compared against the signed Merkle tree root. The same platform and language were used for profiling rendering time of the App and Button components in React DevTools. To evaluate overhead from Merkle-proof verification in the client environment, Web Crypto API was used to compute SHA-256 hashes. The experiment processed 100 synthetic JSON blocks with a Merkle-proof depth of 5, for which the average integrity-verification time was measured. These blocks were selected to provide a representative sample sufficient for assessing algorithm performance under conditions close to real SDUI use. A test module was implemented to assess the interface's protective potential, analysing the number of attacks and the proportion of successful blocking, and presenting the results of PenTest scanning.

Rendering delays of UI components before and after optimisation were visualised under various network-throttling conditions. The graph was built within a canvas element using the Chart.js API and automatically initialised after the React component loaded. Synthetic pre- and post-optimisation delay values were specified for four typical networks (normal, slow3G, fast3G, Wi-Fi), represented as two data series. Additionally, cache-hit rate dynamics over time under different load conditions (low, medium, and high) were demonstrated. Cache-hit-rate simulations were performed on the StackBlitz platform using Chart.js, modelling 50 cache requests across 30-time intervals under three load conditions with base hit probabilities and random ± 5% fluctuations to simulate real environments.

For collecting interface-performance metrics in the client environment, the official web-vitals package from Google was used, implementing measurement of key Web Vitals metrics – Time to First Byte (TTFB), First Contentful Paint (FCP), Largest Contentful Paint (LCP), and Cumulative Layout Shift (CLS). Additionally, examples of input JSON requests and output blueprint files were created using the JSON Editor Online platform, along with an implemented UI-component change log in JSON format. The Service Organisation Control 2 (SOC 2) and Open Worldwide Application Security Project (OWASP) were analysed. The technical specifications of the experimental environment included an HP 250 15.6 inch G10 personal computer (Intel Core i5 processor) and Google Chrome browser version 138.0.7204.185.

Results

Architecture of a secure user interface change log

In modern distributed systems, where the user interface is generated on the server side, the authenticity, and immutability of the UI delivered to the client acquire critical importance. This is especially relevant for financial, medical, and governmental applications, where even a minor alteration of interface elements may cause data leakage, fraud, or regulatory non-compliance. To minimise the risks of manipulation at the client level or within the intermediary infrastructure (content delivery network, proxy, edge functions), an architecture with cryptographic verification of changes and subsequent audit capability is required. In response to these challenges, the CVCLog concept is introduced, providing SDUI systems with means to ensure integrity, verifiability, and immutability of interface content.

CVCLog is a formalised data structure that records every change in UI blocks transmitted from the server to the client, with a guarantee of impossibility of undetected editing or deletion. The foundation of the log consists of structured JSON packets describing the state of UI elements (buttons, forms, messages), along with cryptographic signatures and a Merkle tree of hashes ensuring the integrity of the entire sequence of changes. The log may be implemented as a temporary in-memory repository (e.g., Redis), a permanent log in a database (PostgreSQL or Append-Only Log in S3), or as a hybrid with caching of the latest changes. Each request to the interface (for example, /ui/tenantA/admin) returns not only the UI document but also a Merkle proof, allowing the client to independently verify its authenticity. This ensures non-repudiation of changes, compliance with audit requirements, and trust in the UI even within a compromised environment.

To ensure modification resistance of the user interface in SDUI, each interface block is formed as a structured JSON packet describing its appearance, behaviour, and interaction parameters. To guarantee immutability, such a packet is digitally signed on the server side, and the signed packets are organised into a Merkle tree, enabling rapid and efficient verification of the authenticity on the client side. Thus, any attempt to modify an individual interface element results in a disruption of the hash-tree structure, which is immediately detected during verification. Below is a basic example of constructing a JSON block for a button and computing its hash with a digital signature (Fig. 1).

The code shown is written in JS using the SHA-256 algorithm. This example demonstrates how an individual interface element (e.g., a payment confirmation button) can be protected by hashing and signing already at the stage of generation on the server. The computed SHA-256 hash guarantees detection of even the smallest data changes, while the HMAC signature records the source of generation. Such atomic blocks, signed at the moment of creation, serve as the building elements for forming a full CVCLog, which is subsequently used for verifying integrity and immutability on the client side. This enables SDUI systems to detect falsifications or substitutions of elements even in the event of content delivery network or proxy compromise. To better understand the protection mechanisms embedded in the CVCLog architecture, it is necessary to examine the fundamental theoretical components: Merkle trees, the principles of non-repudiation, digital signatures, and basic concepts of access and replication (Table 1).

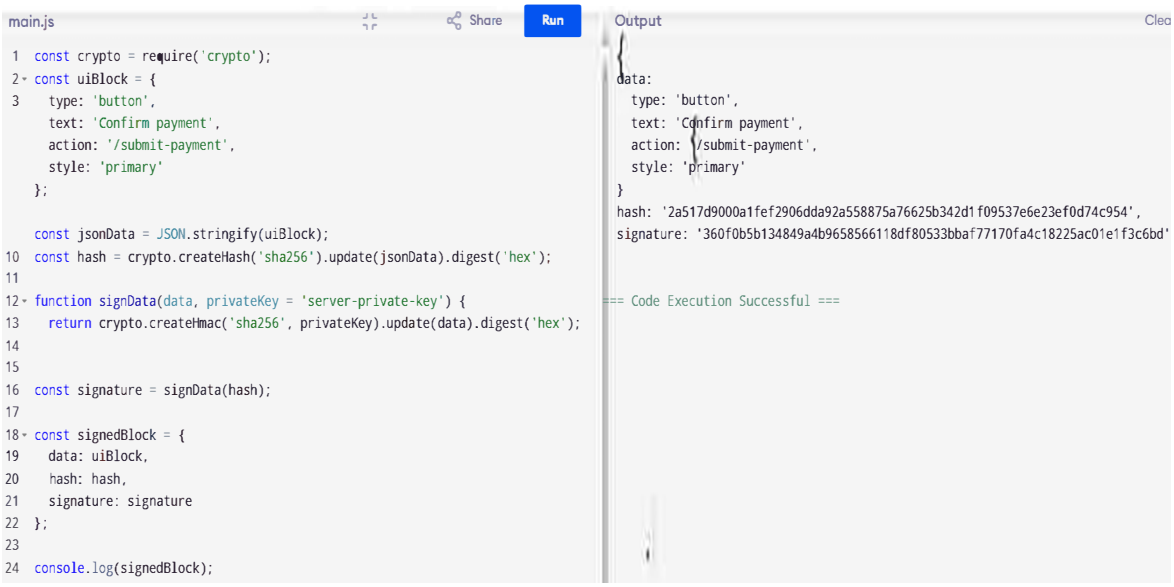


Figure 1. Generation of a signed UI block with hashing and HMAC signature based on JSON
Source: created by the authors

Table 1. Theoretical foundations of ensuring UI content authenticity in CVCLog

Component	Brief description	Example of use in CVCLog/SDUI
Merkle tree	A hierarchical hash structure ensuring integrity and non-repudiation with lower computational cost	Verification of UI JSON-block integrity via root hash and Merkle proof
Non-repudiation principle	Guarantee that the author cannot deny having signed a message, implemented through hashes and Merkle trees	The server cannot deny creation of UI blocks; the proof records the sequence
Digital signature	Cryptographic authentication mechanism using key pairs to protect against forgery	Each UI JSON block is signed by the server; the client verifies the signature
ACL	Access control lists defining user/role permissions to resources	Restricting access to UI blocks (e.g., administrative forms)
CRDT	Data enabling consistent merging of changes in distributed systems even during network failures	Replication of UI states in offline modes or on the edge without loss of consistency

Source: created by the authors based on M. Badra & R. Borghol (2018), X.-Q. Cai *et al*. (2022), P.S. Almeida (2024)

Table 1 summarises the key theoretical components that ensure reliability, authenticity, and immutability of UI content in the CVCLog architecture, highlighting efficient approaches capable of functioning even under limited computational resources or quantum threats. These components underpin the construction of reliable, scalable SDUI systems. In general, SDUI is an architectural approach in which the server defines and generates the user interface, sending the client a description of the UI in the form of structured data, usually JSON. The client, in turn, receives this data and renders the interface according to the provided instructions. This approach enables centralised UI management, ensuring consistency and flexible updates without requiring client-side application releases. Therefore, SDUI emerged as a response to the limitations of traditional architecture, where any interface modification required a new client release, slowing development and deployment of updates.

In this context, the term Backend-Driven UI is often used synonymously with SDUI, but it emphasises not only interface management but also delegation of client logic to the server, including display rules, element behaviour, and adaptation to roles, context, or A/B testing. Here, the backend serves as the single source of truth for the entire interface. It is also worth noting that in modern SDUI implementations, the server not only sends a UI description as JSON but directly streams ready-to-render React components through React Server Components, dynamically forming the UI depending on the response of large language models or real-time data. Such an approach within Backend-Driven UI allows delegating not only structure but also display logic, validation, and behavioural rules to the server, simplifying the client side and ensuring flexible interface adaptation to user context, request, or role. The technological foundation for efficient implementation of the Backend-Driven UI concept has been the development of React – particularly the introduction of server components. A brief timeline of React rendering evolution demonstrates key stages of the technology's progression, from client-side rendering in React 16 through server-side rendering and concurrent rendering to the modern server components in React 19. Figure 2 illustrates these milestones with major innovations and release years.

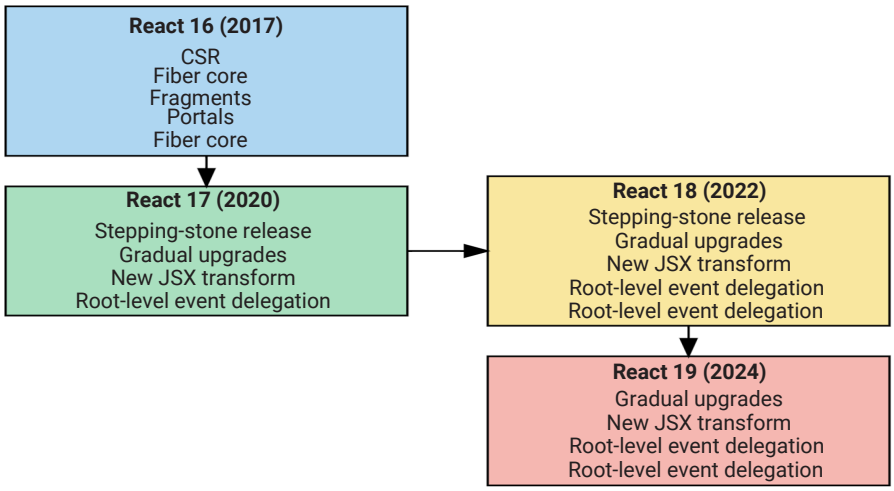


Figure 2. Evolution of React rendering: from React 16 to React 19
Note: CSR – Client-Side Rendering, JSX – JavaScript eXtensible Markup Language, SSR – Server-Side Rendering
Source: created by the authors

The presented scheme demonstrates the gradual sophistication and optimisation of React rendering approaches, enhancing interface performance, flexibility, and scalability. It is on this technological basis that CVCLog integration with the React client is realised. The integration of the log involves implementing mechanisms for verification of Merkle proofs directly on the client side. React components receive signed JSON blocks and corresponding integrity proofs, enabling real-time verification of UI authenticity and prevention of rendering falsified or modified content. This approach provides reliable interface protection even in the event of network-infrastructure compromise.

Given the gradual development of rendering technologies in React, integration of CVCLog with the client side becomes an important step in ensuring interface integrity and security. However, to implement such mechanisms effectively, compliance with regulatory requirements concerning data protection and transaction security must also be considered. Table 2 summarises the key standards influencing the architecture of the secure change log.

Table 2. Regulatory compliance of CVCLog architecture

Standard	Brief description	Impact on CVCLog architecture
PCI-DSS	Global security standard for organisations handling payment cards, emphasising endpoint and access-log protection	Ensures UI integrity through cryptographic signatures and Merkle trees; meets log-protection and client-security requirements
PSD2	EU regulation for payment services requiring strong authentication and access control	Guarantees a secure and transparent UI, minimising forgery and unauthorised-access risks
GDPR	EU regulation for personal-data protection emphasising privacy, encryption, and audit	Minimises processing of clients' personal data, ensures transparency and non-repudiation of changes, complies with security and access-control requirements

Source: created by the authors based on M.K. Agarwal *et al.* (2024), L.D. Christensen (2025), H. Sienkiewicz (2025)

These standards define requirements for protection, integrity, and transparency of interface content in the CVCLog architecture, underscoring the importance of cryptographic verification and access control for compliance with modern security norms. Meanwhile, disruption of the UI flow – illustrating insufficient interface protection – may have serious consequences, as evidenced by specific incident examples from various countries. For instance, in September 2024, a vulnerability in 7-Zip allowed attackers to bypass Windows Mark-of-the-Web protection, leading to execution of malicious code via double archiving (Girnus, 2025). This vulnerability was actively exploited in the SmokeLoader campaign targeting governmental and public organisations in Ukraine. In the United States, in 2025 the medical company Episource suffered a data breach affecting 5.4 million users due to server compromise (Fadilpašić, 2025). Attackers gained access to medical records, personal data, and insurance information.

To summarise, the CVCLog architecture is based on cryptographic verification of signed JSON blocks organised into a Merkle tree, ensuring immutability and non-repudiation of UI changes even in complex distributed systems. This approach integrates with modern SDUI and React technologies, allowing the client to verify UI-content authenticity in real-time while complying with strict regulatory requirements (PCI-DSS, PSD2, GDPR). Thus, CVCLog provides a robust protective layer against UI tampering, reducing the risks of data leakage, fraud, and security-breach penalties.

Implementation of verification mechanisms, API audit and experimental evaluation

To prevent the visualisation of compromised interface blocks in the client environment, implementation of a mechanism for verifying JSON-structure integrity using Merkle proofs is crucial. The first stage involves creating a client function that validates a Merkle proof based on the input JSON, proof tree, and signed root. Only after successful verification are the data passed to the React component. Figure 3 shows a basic implementation of this mechanism.

```jsx
import React, { useEffect, useState } from 'react';
import sha256 from 'crypto-js/sha256';

const verifyProof = (leaf, proof, root) => {
  let hash = sha256(leaf).toString();
  for (const { hash: sibling, position } of proof) {
    hash =
      position === 'left'
        ? sha256(sibling + hash).toString()
        : sha256(hash + sibling).toString();
  }
  return hash === root;
};

export default function App() {
  const [isValid, setIsValid] = useState(null);
  const [duration, setDuration] = useState(null);

  const jsonBlock = JSON.stringify({
    component: 'Button',
    props: { text: 'Submit', style: 'primary' },
  });

  const proof = [
    { hash: sha256('sibling1').toString(), position: 'left' },
    { hash: sha256('sibling2').toString(), position: 'right' },
  ];
```

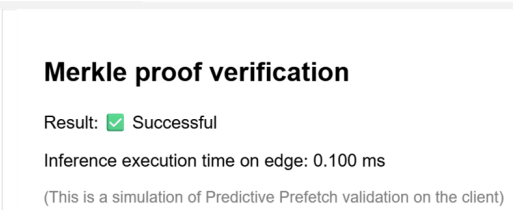


Figure 3. Fragment of client verification code of a Merkle proof for a JSON block in React
Source: created by the authors

In this program, an example of a UI component of the Button type is demonstrated, which is serialised into JSON and hashed for further integrity verification. The computation is performed by sequentially combining the hashes of the JSON block with the hashes of the proof elements according to the positions, after which the resulting value is compared with the Merkle tree root hash. In the given case, the verification passes successfully, which is reflected in the status of the React component. Additionally, the execution time of the operation on the client side is measured – 0.1 milliseconds (ms) – simulating the performance of the Predictive Prefetch mechanism. This approach allows integrity verification to be integrated directly into the client part of SDUI, ensuring protection from forgery or unauthorised modifications of interface elements before the rendering.

The next step is the implementation of reactive verification, which enables real-time tracking of changes in interface data and automatic integrity checks during updates. Unlike one-time verification before rendering, reactive verification provides continuous monitoring, which is particularly important in dynamic SDUI where changes occur constantly. It is based on observing data streams or state-change events, which makes it possible to respond promptly to potential integrity violations.

In general, for the effective integration of SDUI mechanisms into React applications, it is important to ensure the possibility of gradual loading and activation of interface components. This functionality is implemented by React's architectural capabilities – Suspense, Streaming, and Partial Hydration. Suspense makes it possible to "pause" rendering of individual components until asynchronous operations, such as obtaining a blueprint configuration from the server, are completed. In combination with server streaming, this allows the interface to be delivered in parts as soon as the components are ready, reducing the time to first render and increasing overall performance.

In turn, Partial Hydration enables activation of only those DOM parts that have dynamic behaviour, leaving static elements uninitialised until necessary. This allows flexible integration of SDUI architecture with React without complete re-initialisation of all components on the client side. The next stage is the implementation of an API-call scenario for obtaining blueprint configurations depending on the user context (Fig. 4).

```json
{
  "tenant": "acme-manager",
  "role": "admin"
}
```

Figure 4. Input JSON request
Source: created by the authors

In this example, the request is sent to the endpoint GET/ui/{tenant}/{role} with corresponding parameters. It identifies the user as an administrator of the organisation, based on which the server generates the appropriate blueprint configuration of the interface. In response, the server

sends a blueprint file – a JSON structure containing a description of interface components together with metadata required for client-side integrity verification (Fig. 5).

```
{
  "components": [
    {
      "type": "Button",
      "props": {
        "label": "Create Report",
        "action": "/reports/create"
      },
      "hash": "fae124d8d9f7d3b2c5e87cb309e46c098c",
      "proof": [
        { "position": "left", "hash": "a1b2c3d4e5f6..." },
        { "position": "right", "hash": "d4e5f6a1b2c3..." }
      ]
    }
  ],
  "merkleRoot": "3c6e0b8a9c15224a8228b9a98ca1531d"
}
```

Figure 5. Output blueprint file
Source: created by the authors

In this example, the server returns a Button component with the label Create Report and the action /reports/create. The response also contains the hash of each block, the Merkle proof, and the signed root hash. Such a structure enables independent client-side verification of the received data before the rendering, which is a key principle of secure SDUI. To ensure traceability of changes in SDUI, it is necessary to implement API auditing, which records every modification of blueprint configuration – including content, structure, or access-rights changes to UI components. Such auditing is critically important in the context of access control to sensitive interface elements (i.e., PII zones) according to the requirements of SOC 2 and OWASP. For example, SOC 2 is an audit standard defining requirements for control over security, confidentiality, processing integrity, availability, and data privacy (Joodala, 2025). It requires active logging of access to UI components with confidential or personal data, control of interface-configuration changes, and the ability to reproduce the history of operations on PII zones for auditing purposes.

In turn, OWASP Top 10 is an international standard identifying the most critical web-application vulnerabilities (Li & Li, 2025). It is based on the analysis of thousands of incidents and covers threats related to data integrity, access control, API protection, and security configuration. The 2025 version of OWASP Top 10 focuses on risks associated with artificial intelligence (AI), complexity of APIs and cloud systems, software-supply-chain attacks, and the strengthening of regulatory requirements. Compliance with these recommendations is essential for ensuring SDUI security, particularly through logging, integrity verification, and access control. To implement such control, detailed logging of every interface-configuration change must be maintained, ensuring transparency and the ability to restore the change history at any time. Each time a UI file is updated on the server, the event is recorded in a change log considering the time of change, author, UI-fragment identifier, hash of the new state, and reference to the MerkleRoot. Figure 6 shows an example of a change-log entry.

```
{
  "timestamp": "2025-07-23T14:52:11Z",
  "tenant": "acme-manager",
  "role": "admin",
  "modifiedBy": "user_42",
  "changeType": "update",
  "componentId": "button_create_report",
  "newHash": "e7f8a9d1c2b3...",
  "merkleRoot": "3c6e0b8a9c15224a8228b9a98ca1531d"
}
```

Figure 6. Fragment of the UI component change log in JSON format
Source: created by the authors

In addition to logging, a mechanism for restoring the history of changes is implemented: using the API request GET /ui/history/{tenant}/{role}, previous versions of blueprint files with the hashes and timestamps can be retrieved. This allows not only forensic analysis but also restoration of a previous state in case of compromise. Such an approach to API auditing integrates integrity, observability, and rollback of changes into a single secure UI-management cycle that complies with the principles of transparency and accountability in modern information systems. For timely notification of clients about changes in UI configuration and to ensure cache validity, the Redis publish-subscribe (Pub/Sub) mechanism is used (Fig. 7).

```
[2025-07-24T10:15:32Z] Channel: ui_invalidation
Message: {
  "tenant": "acme-manager",
  "role": "admin",
  "updatedComponent": "Button",
  "newMerkleRoot": "3c6e0b8a9c15224a8228b9a98ca1531d",
  "timestamp": "2025-07-24T10:15:32Z"
}
```

Figure 7. Example of Redis Pub/Sub message about cache invalidation
Source: created by the authors

When a blueprint file is updated, the server publishes a message in a Redis channel, signalling the need to invalidate the cache of the corresponding UI block. Figure 7 illustrates an example of a Redis message about cache invalidation of a UI component. In the example provided, the server notifies about the update of the Button component for the administrator, indicating the new merkleRoot value, allowing clients to promptly delete outdated cache and request current data. To analyse rendering overhead after receiving updates, the React DevTools Profiler instrument was used (Fig. 8).

Figure 8. Profiling of rendering time of App and Button components in React DevTools
Source: created by the authors

The results showed that rendering of the App component lasted 2.7 ms, of which 1.8 ms was active updating, while the Button component was rendered for 0.4 ms (of which 0.2 ms was actual change). This indicates low overhead during UI updates in response to configuration changes, confirming the feasibility of integrating client-side verification mechanisms in dynamic SDUI applications. A complement to this analysis is the data obtained during performance modelling of client-side verification and evaluation of interface-level attack resistance, presented in Figure 9.

```
import React, { useEffect, useState } from 'react';

function hashJsonBlock(block) {
  return crypto.subtle
    .digest('SHA-256', new TextEncoder().encode(block))
    .then((buf) =>
      Array.from(new Uint8Array(buf))
        .map((b) => b.toString(16).padStart(2, '0'))
        .join('')
    );
}

async function verifyMerkleProof(leafHash, proof, root) {
  let current = leafHash;
  for (const sibling of proof) {
    const combined = current + sibling;
    current = await hashJsonBlock(combined);
  }
  return current === root;
}

async function simulateExperiment(numBlocks = 100, proofDepth = 5) {
  let totalTime = 0;
  let pass = 0;
```

Experiment: Overhead + PenTest

✅ **Merkle Proof Verification**

- Number of blocks: 100
- Successfully tested: 92
- Average time: 0.328 ms

🔐 **PenTest Simulation**

- Attack scenarios: 50
- Blocked: 47
- Block rate: 94.0 %

Figure 9. Fragment of React application for evaluating Merkle-proof verification overhead and results of PenTest scanning
Source: created by the authors

This program demonstrates an experimental evaluation of overhead for verifying Merkle proofs of 100 JSON-data blocks, of which 92 (92%) were successfully verified, with an average verification time of about 0.328 ms per block, indicating the algorithm's efficiency in the client environment. Also presented are the results of simulated PenTest scanning, in which 47 of 50 attack scenarios were successfully blocked, corresponding to a 94% protection level, demonstrating the potential of interface-level defence mechanisms. To better assess the effect of optimisations on interface performance, a separate analysis of UI-component rendering delays under various network conditions using network throttling was conducted, the results of which are shown in Figure 10.

The graph compares delay before and after optimisation in the client React application under different network conditions. The highest delay before optimisation was observed under slow3G – up to 850 ms – whereas after optimisation it decreased to 300 ms. Under Wi-Fi and normal-latency conditions, the delay dropped to less than 50 ms, demonstrating a significant reduction in rendering time with low network latency. The graph illustrates that optimisation mechanisms (e.g., predictive preloading, partial hydration) effectively reduce latency across all connection types, which is particularly critical for dynamic SDUI applications. However, it is equally important to maintain a high cache-reuse ratio, especially under different load conditions – this dependence is shown in Figure 11.

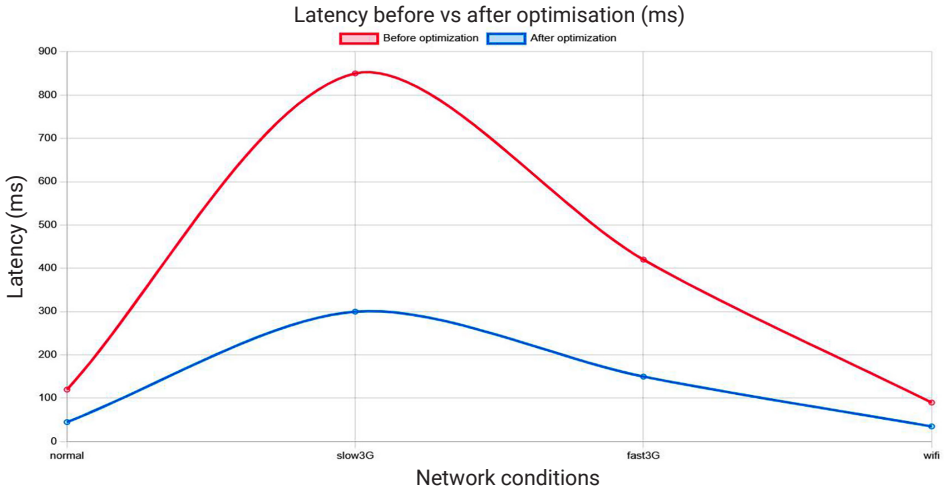


Figure 10. Graph of UI-component rendering delay before and after optimisation under various network-throttling conditions

Source: created by the authors

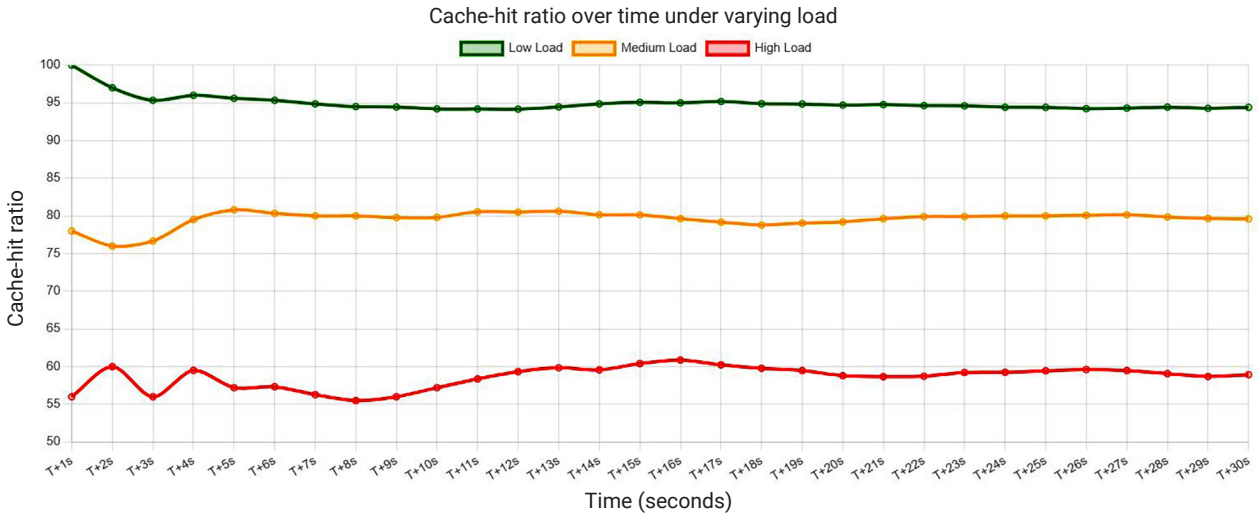


Figure 11. Graph of cache-hit rate over time under different load conditions

Source: created by the authors

The graph displays cache-hit-rate changes over a 30-second period under three load conditions (low, medium, and high), with a base hit probability defined for each (0.95, 0.80, and 0.60, respectively). Low load consistently shows the highest hit rate – from 100% initially to stable 94-95% after a few seconds. Medium load exhibits a lower level – around 77-81%, tending to stabilise near 80%. High load records the lowest cache-hit rate – between 55% and 61% – with notable variability at the start and stabilisation around 58%. Thus, the results demonstrate a clear dependence of caching efficiency on load: as load increases, the hit rate decreases, which is critically important for SDUI systems with high-frequency interface updates.

Additionally, Web Vitals are important – standardised metrics reflecting real user experience when interacting with an interface. For example, TFB is the time between a request and receipt of the first byte of the response; FCP is the time until the browser displays the first visual DOM element; LCP is the rendering time of the largest visible element on the screen. For analytics, quantiles p50 (median) and p95 (worst 5% scenario) are used. The measured values before and after optimisation are shown in Figure 12.

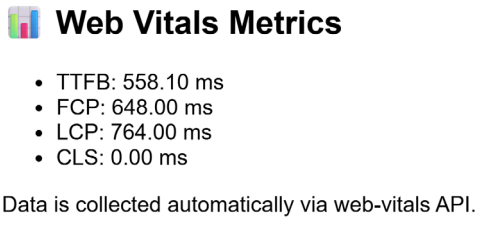


Figure 12. Measurement of Web Vitals metrics before/after cache and Predictive Prefetch implementation

Source: created by the authors

After implementing caching of blueprint configurations and the Predictive Prefetch algorithm, improvements in FCP and LCP were observed due to reduced network-loading delays. Specifically, p95 LCP metrics decreased by tens of milliseconds, which is particularly noticeable during repeated component loads. The results confirm the expediency of integrating predictive preloading into SDUI applications to improve performance.

Thus, implementation of interface-integrity verification based on Merkle proofs ensures tamper-resistance of SDUI applications, preventing undetected modifications of interface blocks even in a compromised environment. The introduced optimisations – including caching, Predictive Prefetch, and partial hydration – significantly reduce rendering delays and improve key Web Vitals metrics, especially under limited network resources. A comprehensive approach combining client-side proof verification, API auditing, cache control, and reactive verification forms a robust architecture that meets regulatory-standard requirements and ensures a high level of security, transparency, and resistance to interface-level attacks.

Discussion

In this work, client verification of Merkle proofs in SDUI was implemented, ensuring protection of the interface from unauthorised modifications with a verification delay of 0.328 ms, even under high load conditions. In turn, M. Ethan (2025) proposed a frontend-driven backpressure handling model in real APIs, which improves latency, memory control, and interface stability under load. Hence, both approaches aim to ensure the resilience of the client side to overload, but the current system additionally guarantees cryptographic data integrity.

In the conducted study, an SDUI model was implemented with client-side integrity verification of signed JSON packages using Merkle proofs, which prevents interface configuration tampering and ensures immutability similar to blockchain solutions. Conversely, in the work of I. Shahzad *et al.* (2025), an Internet of Things (IoT) system using a trusted blockchain for secure sensor management and data recording with immutability and controlled access guarantees was presented. Both approaches are aimed at enhancing trust in dynamic systems through the use of immutable data structures; however, the current solution focuses on verifying interface data on the client side before rendering.

Within this study, a mechanism for verifying interface JSON structures in SDUI through Merkle proofs was created, ensuring auditability of dynamic UI without loss of performance. On the other hand, G. Sharma (2025) proposed an architecture with a blockchain embedded in the operating system kernel for immutable logging of AI decisions, achieving 100% accuracy in forgery detection and reducing audit preparation time to real-time. Both approaches align in the pursuit to ensure transparency and compliance with audit requirements through record immutability, although the current solution achieves this at the client-interface level with minimal delays, whereas the mentioned model implements similar properties at the OS-kernel level, accompanied by higher system costs.

The results proposed the CVCLog model for SDUI, ensuring immutability, authenticity, and detailed auditing of UI components in real-time without performance loss. In comparison, S. Fugkeaw *et al.* (2025) presented the scheme Efficient and Verifiable Searchable Encryption with Boolean Search, which ensures integrity and searchable accessibility of cloud logs through a combination of indexing, blockchain ACL, and Merkle-root verification. Both approaches are consistent in striving to guarantee controlled immutability of data; however, the current system is oriented specifically at the UI level and allows detection of modifications even before rendering, which is critical in a real-time context.

Analysis of the requirements of PCI-DSS, PSD2, GDPR, SOC2, and OWASP standards for interface integrity, change logs, and security of client components enabled the formation of a cryptographically verifiable UI architecture that complies with regulatory requirements. In turn, P.R. Vennamaneni (2025) examined the compliance of AI systems in the financial sector with PCI-DSS and IEC 62304 standards, focusing on transaction data protection, decision-making transparency, and cloud infrastructure. Thus, the results of the presented study confirm the expediency of adhering to standards in critical digital systems, while the current work demonstrates a concrete implementation of this compliance at the client-interface level.

This work examined the evolution of React rendering (versions 16-18), integration of Server Components, Suspense, and Partial Hydration to ensure continuous rendering during client verification of Merkle proofs, enabling UI integrity control prior to its display. In turn, S. Wagh *et al.* (2025) presented React-Nex – a modular component library with AI-driven code generation support via Retrieval-Augmented Generation and vector embeddings, accelerating the creation and configuration of interface elements. Both approaches are aimed at improving the efficiency of React-application development; however, the current solution focuses on interface protection and verification, whereas React-Nex is oriented towards automation and flexibility in UI-component generation.

The proposed CVCLog architecture for SDUI ensures cryptographic integrity of interface data and detailed real-time audit of changes, including through client-side verification of Merkle proofs prior to rendering. Regarding the work of N. Jose (2025), it demonstrated the advantages of an event-driven architecture for real-time synchronisation of inventory data in retail systems, emphasising consistency and event-processing speed. Both approaches align in the pursuit of building reactive, scalable, and transparent real-time systems; however, the current CVCLog architecture additionally addresses the task of verifying data authenticity at the client level, which is critical for secure interactions in distributed interfaces.

The conducted study proposed the CVCLog architecture for SDUI, in which configuration JSON interface blocks

are organised as a Merkle tree, and integrity verification is performed client-side before rendering by computing and comparing Merkle proofs with the signed root hash. Similarly, A. Odeh & A. Abu Taleb (2025) proposed a Blockchain-Enhanced Trust and Access Control for IoT Security model, which uses Merkle trees, blockchain, and federated learning for decentralised access control and data integrity verification in IoT environments. Both approaches aim to strengthen trust in dynamic, distributed systems through built-in authenticity verification and decentralised control mechanisms, although the current solution implements these properties directly in the UI layer, enabling interactive protection of the user interface without performance loss.

Additionally, A. Osilaja *et al.* (2024) justified the use of blockchain for building secure and transparent software architectures, focusing on data immutability, decentralised access control, and cryptographic verification. In alignment with this approach, the current CVCLog architecture is also oriented towards ensuring trust in dynamic systems but implements these principles directly at the user-interface level. The distinction lies in the fact that the proposed solution enables interactive verification of UI-component authenticity in real-time, whereas the mentioned authors focus on general aspects of software security.

The current CVCLog architecture in the SDUI environment ensured verification of interface JSON-data authenticity before rendering, with an average Merkle-proof verification time of 0.328 ms and attack-detection accuracy of up to 94%, without affecting interface performance. Meanwhile, A.A. Sathio *et al.* (2025) proposed the ClusterPioneer model for trusted blockchain, which reduces communication load by 60%, provides verification within 150 ms, and achieves 95% accuracy in attacker detection. Hence, the results of the mentioned work confirm the expediency of using Merkle structures and decentralised verification to ensure data integrity in critical digital systems, while the current solution demonstrates the application of these principles at the client UI level with minimal delays.

The conducted experiments confirmed that integration of cryptographic verification with React Suspense and Partial Hydration ensures stable UI rendering with minimal impact on response time (2.7 ms for the main component) even under high loads. Conversely, Y. Chavan *et al.* (2025) presented Nexify – a scalable real-time server for online communities providing low latency, modularity, role-based access control, and end-to-end encryption, as well as incorporating blockchain identification and AI moderation. Thus, both approaches are oriented towards building secure, scalable, and reactive systems, while the current solution complements the Nexify framework with the ability to verify interface authenticity prior to its display – a critical element of trust in dynamic environments.

Overall, the conducted study focuses on cryptographic verification of UI-data integrity in real-time with minimal delays and high change-detection accuracy at the client level. In turn, the work of B. Ganji *et al.* (2024) focused on formal verification of distributed streaming systems to prevent process deadlocks and errors already at the design stage. Therefore, both approaches complement each other, combining data protection with architecture correctness assurance in complex real-world systems.

To summarise, this work proposed the CVCLog architecture for SDUI, which implements client-side verification of interface JSON-data integrity using Merkle proofs with low latency and high change-detection accuracy. This approach ensures UI protection from unauthorised modifications before rendering, supports detailed real-time auditing, and complies with leading security and audit standards. Thus, the study not only complements existing solutions for load control and data security but also advances practical methods for ensuring trust in dynamic client interfaces in distributed web systems.

Conclusions

The conducted study confirmed the effectiveness of the CVCLog architecture as a reliable mechanism for ensuring integrity and immutability of UI in SDUI systems. The obtained results emphasised that the use of cryptographically signed JSON blocks and Merkle trees allows detection of any unauthorised changes on the client side even under network infrastructure compromise conditions. Implementation of a change log with non-repudiation support guarantees compliance with high regulatory standards such as PCI-DSS, PSD2, and GDPR. Furthermore, integration with the React client ensures verification of UI authenticity in real-time without significant impact on performance.

It was also established that client verification of Merkle proofs in a React application is performed with an average time of 0.1 ms per block, confirming its efficiency for integration into dynamic SDUI. Reactive verification ensures continuous control of UI changes in real-time, reducing the risks of content tampering. Rendering profiling showed that updating the main App component takes 2.7 ms, while an individual Button component requires 0.4 ms, indicating low overhead for client-side verification. In turn, experimental verification of JSON blocks revealed 92% successful verifications, with an average time of 0.328 ms per block. Meanwhile, PenTest scanning results demonstrated 94% attack blocking at the UI level, and cache hit rate in stable mode reached 94-95% under low load and decreased to 58% under high load. Implementation of optimisations improved key Web Vitals metrics: p95 LCP decreased by tens of milliseconds, enhancing response time during repeat loads.

The study's limitations lie in the dependency of verification and caching performance on client hardware resources and network quality, which may restrict the applicability of the proposed mechanisms in low-power or highly constrained environments, as well as in the complexity of scaling reactive verification under very high UI update frequency. For further development, it is recommended to optimise verification algorithms using hardware acceleration and to design adaptive caching strategies considering load variability and network conditions. An important direction is integration of machine-learning mechanisms for

predicting UI changes and preloading the most probable components, which would further reduce rendering latency.

Acknowledgements
None.

Funding
The study was not funded.

Conflict of Interest
None.

References
[1] Agarwal, M.K., Sarden, D., Ramesh, S., & Singh, R. (2024). Endpoint controls through a lens of PCI DSS. In M. Gupta, R. Singh, J. Walp & R. Sharman (Eds.), *Advances in enterprise technology risk assessment* (pp. 245-282). London: IGI Global. doi: 10.4018/979-8-3693-4211-4.ch009.

[2] Almeida, P.S. (2024). Approaches to conflict-free replicated data types. *ACM Computing Surveys*, 57(2), article number 51. doi: 10.1145/3695249.

[3] Azhar, H.B., Butt, K.K., Awan, N.U., & Irshad, O. (2025). Quantum-resistant merkle trees enhancing data integrity with post-quantum cryptography and zero-knowledge proof. *Journal of Computing & Biomedical Informatics*, 8(2). doi: 10.56979/802/2025.

[4] Badra, M., & Borghol, R. (2018). Long-term integrity and non-repudiation protocol for multiple entities. *Sustainable Cities and Society*, 40, 189-193. doi: 10.1016/j.scs.2017.11.023.

[5] Cai, X.-Q., Wang, T.-Y., Wei, C.-Y., & Gao, F. (2022). Cryptanalysis of quantum digital signature for the access control of sensitive data. *Physica A: Statistical Mechanics and its Applications*, 593, article number 126949. doi: 10.1016/j.physa.2022.126949.

[6] Chavan, Y., Jadhav, A., Kulkarni, S., Malpure, S., & Mandal, S. (2025). Nexify: A scalable and secure community server for real-time communication. *International Journal of Advanced Research in Science Communication and Technology*, 5(4), 547-551. doi: 10.48175/IJARSCT-25172.

[7] Christensen, L.D. (2025). Financial fraud and the PSD2. In L.D. Christensen (Ed.), *EU payment services: Regulation and innovation* (pp. 145-181). Oxford: Oxford University Press. doi: 10.1093/9780198949084.003.0006.

[8] Du, P., Liu, Y., Li, Y., & Yin, H. (2022). EthMB+: A tamper-proof data query model based on b+ tree and Merkle tree. In Y. Sun, L. Cai, W. Wang, X. Song & Z. Lu (Eds.), *Blockchain technology and application* (pp. 49-59). Singapore: Springer. doi: 10.1007/978-981-19-8877-6_4.

[9] Ethan, M. (2025). *Frontend-driven backpressure handling for real-time APIs*. Retrieved from https://www.researchgate.net/publication/393981918_Frontend-Driven_Backpressure_Handling_for_Real-Time_APIs.

[10] Fadilpašić, S. (2025). *Major breach at medical billing giant sees data on 5.4 million users stolen – here's what we know*. Retrieved from https://www.techradar.com/pro/security/major-breach-at-medical-billing-giant-sees-data-on-5-4-million-users-stolen.

[11] Fugkeaw, S., Deevijit, J., Ueasathitwong, P., & Thanyasukpaisal, T. (2025). EVSEB: Efficient and verifiable searchable encryption with boolean search for encrypted cloud logs. *IEEE Access*, 99, 101177-101195. doi: 10.1109/ACCESS.2025.3577466.

[12] Ganji, B., Rezaee, A., Adabi, S., & Movaghar, A. (2024). Model verification of real-time and distributed stream processing architecture. *Computing*, 107(1), article number 17. doi: 10.1007/s00607-024-01384-w.

[13] Girnus, P. (2025). *CVE-2025-0411: Ukrainian organizations targeted in zero-day campaign and homoglyph attacks*. Retrieved from https://www.trendmicro.com/en_us/research/25/a/cve-2025-0411-ukrainian-organizations-targeted.html.

[14] Havatiuk, M., & Saiapina, I. (2025). Improved method of targeted user interface updates for enhancing the efficiency of web applications based on reactive streams and virtual DOM. *Technical Engineering*, 95(1), 259-265. doi: 10.26642/ten-2025-1(95)-259-265.

[15] Joodala, A. (2025). A cloud-native approach to SOC 2, HIPAA, and GDPR compliance using AWS microservices. *International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences*, 13(3). doi: 10.37082/IJIRMPS.v13.i3.232605.

[16] Jose, N. (2025). Event-driven architecture in retail: Real-time inventory synchronization for omnichannel retail. *International Journal of Computing and Engineering*, 7(16), 13-23. doi: 10.47941/ijce.3014.

[17] 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.

[18] Li, J., & Li, H. (2025). Evolution of application security based on OWASP top 10 and CWE/SANS top 25 with predictions for the 2025 OWASP top 10. In *8th International conference on inventive computation technologies* (pp. 1178-1183). Kirtipur: IEEE. doi: 10.1109/ICICT64420.2025.11004742.

[19] Odeh, A., & Abu Taleb, A. (2025). Federated learning and blockchain framework for scalable and secure IoT access control. *Computers, Materials & Continua*, 84(1), 447-461. doi: 10.32604/cmc.2025.065426.

[20] Orynchak, A., Kuzmenko, O., & Svintsytska, O. (2024). Real-time threat detection with javascript: Monitoring and response mechanisms. *Technical Engineering*, 93(1), 201-210. doi: 10.26642/ten-2024-1(93)-201-210.
[21] Osilaja, A., Raheem, A., & Edmund, E. (2024). Enhancing software security with blockchain integration for decentralized and tamper-proof application architectures. *World Journal of Advanced Research and Reviews*, 24(3), 2750-2767. doi: 10.30574/wjarr.2024.24.3.3977.
[22] Patel, O. (2022). Merkle proof verification for zero knowledge transaction validation. *International Journal of All Research Education & Scientific Methods*, 10(5), 3533-3547.
[23] Ridhorkar, S., & Mishra, S.S. (2024). Implementing quantum resistant algorithm in blockchain-based applications. *International Journal of Advanced Research in Science Communication and Technology*, 4(7), 650-659. doi: 10.48175/IJARSCT-17899.
[24] Rubel, M.T., Emran, A.K., Islam, M.K., Nayem, M.A., & Hasan, S. (2025). From ledger to ledgerless: Evaluating blockchain-driven real-time financial reconciliation in U.S. public companies. *International Journal for Multidisciplinary Research*, 7(4). doi: 10.36948/ijfmr.2025.v07i04.49709.
[25] Sathio, A.A., Rind, M.M., & Awan, S.A. (2025). ClusterPioneer voting: A scalable and energy-efficient consensus mechanism for permissioned-blockchain (DeFi) system. Research Square. doi: 10.21203/rs.3.rs-7099560/v1.
[26] Shahzad, I., Maqsood, M.W., Latif, S., & Ijaz, H.M. (2025). Decentralized IoT-based architectures for tamper-proof agricultural sensor networks: Ensuring end-to-end data integrity and transparent governance. *Kashf Journal of Multidisciplinary Research*, 2(5), 39-55. doi: 10.71146/kjmr442.
[27] Sharma, G. (2025). Kernel-embedded blockchain architecture for transparent AI decision auditing. *Journal of Information Systems Engineering & Management*, 10(47), 183-205. doi: 10.52783/jisem.v10i47s.9246.
[28] Shport, L. (2025). Enhancing the security of interbank payments with, a comprehensive cryptographic architecture. *Information Technology and Society*, 16(1), 276-280. doi: 10.32689/maup.it.2025.1.36.
[29] Sienkiewicz, H. (2025). *Article cybersecurity impacts of the EU GDPR*. Retrieved from https://www.researchgate.net/publication/393802678_Article_Cybersecurity_Impacts_of_the_EU_GDPR.
[30] Undirwadkar, A.J. (2025). The rise of server-driven UI: A paradigm shift in mobile app development. *World Journal of Advanced Engineering Technology and Sciences*, 15(2), 55-61. doi: 10.30574/wjaets.2025.15.2.0538.
[31] Vennamaneni, P.R. (2025). Building compliance-driven AI systems: Navigating IEC 62304 and PCI-DSS constraints. *International Journal of Networks and Security*, 5(1), 62-90. doi: 10.55640/ijns-05-01-06.
[32] Wagh, S., Vadhel, S., Tiwari, R., Bidaye, V., & Kachare, A. (2025). React-Nex – a modular component library with AI-driven code generation. *International Journal of Scientific Research in Engineering and Management*, 9(4), 1-9. doi: 10.55041/IJSREM44477.

Тампер-резистентна архітектура Server-Driven UI з верифікацією Merkle-доказів у реальному часі

Владислав Ананченко

Аспірант
Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука
33000, вул. Степана Дем'янчука, 4, м. Рівне, Україна
https://orcid.org/0009-0004-8963-775X

Юрій Лотюк

Кандидат педагогічних наук, доцент
Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука
33000, вул. Степана Дем'янчука, 4, м. Рівне, Україна
https://orcid.org/0000-0001-6696-5583

Анотація. Системи інтерфейсу користувача, що керуються сервером, потребують захисту від несанкціонованих змін для забезпечення цілісності і безпеки даних, що відображаються. Мета цього дослідження полягала у розробці криптографічно перевірного журналу змін інтерфейсу користувача для систем із Server-Driven User Interface. У межах дослідження застосовано методи теоретичного моделювання, експериментального тестування, програмної реалізації та аналізу нормативної бази для розробки, верифікації та оцінки криптографічного журналу змін у середовищі клієнтського інтерфейсу. Основні результати показали, що застосування підписаних структурованих блоків інтерфейсу із хешуванням і цифровим підписом забезпечує неможливість непомітної модифікації інтерфейсу на клієнтській стороні. Побудова журналу змін на основі дерева хешів гарантує достовірність, незмінність і криптографічну перевірку кожного елементу інтерфейсу навіть у складних розподілених умовах. Інтеграція з новітніми механізмами рендерингу React дозволяє здійснювати перевірку достовірності інтерфейсу в реальному часі, забезпечуючи відповідність вимогам міжнародних стандартів захисту персональних даних і безпеки транзакцій. Крім того, результати показали, що клієнтська перевірка Merkle-доказів для блоків у React дозволяє виявити модифікації до моменту рендерингу, із середнім часом верифікації 0,328 мілісекунди на блок. Аудит змін blueprint-файлів і система публікації-підписки забезпечили відстежуваність і актуальність даних, тоді як рендеринг компонентів після оновлень тривав лише 2,7 мілісекунди для основного компонента і 0,4 мілісекунди для кнопки. Експерименти підтвердили досягнення 94 % рівня блокування атак, зниження затримок рендерингу (з 850 до 300 мілісекунд у повільній мережі) та підтримку кеш-хітрейту на рівні 94–95 % при низькому навантаженні, що разом із покращенням ключових показників взаємодії з інтерфейсом демонструє ефективність запропонованої архітектури. Отримані результати можуть бути використані розробниками критичних вебзастосунків для впровадження захищених інтерфейсів, що перевіряють цілісність у реальному часі та відповідають міжнародним вимогам безпеки

Ключові слова: криптографічно перевірний журнал змін; принцип незаперечності; еволюція React; оцінка накладних витрат; кеш-хітрейт

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 04.09.2025 Revised: 18.11.2025 Accepted: 23.12.2025

UDC 004.03 DOI: 10.31649/vitce/3.2025.23

Three-scenario analysis of fault diagnosis accuracy in complex technical systems

Vladimir Vychuzhanin

Doctor of Technical Sciences, Professor
Odesa Polytechnic National University
65044, 1 Shevchenko Ave., Odesa, Ukraine
https://orcid.org/0000-0002-6302-1832

Alexey Vychuzhanin

PhD, Assistant
Odesa Polytechnic National University
65044, 1 Shevchenko Ave., Odesa, Ukraine
https://orcid.org/0000-0001-8779-2503

Abstract. The purpose of this study was to perform a three-scenario comparative analysis of the accuracy of intelligent fault diagnosis in complex technical systems using ship power plants (SPPs) as a representative case. The research sought to determine which diagnostic configuration ensures the highest accuracy and robustness under varying operational conditions. Three methodological configurations were analysed: a baseline model based on Case-Based Reasoning (CBR); CBR enhanced with probabilistic analysis using Bayesian networks and Markov chains; and a comprehensive integration of CBR with probabilistic models and simulation modelling of cascading failures. Experiments were conducted under three typical operational scenarios – nominal mode, high-load mode, and limited diagnostic data – reflecting real maritime conditions. Standard classification accuracy metrics were employed, including Accuracy, Recall, and F1-score. The results showed that the basic CBR configuration achieves an average accuracy of 82-85% under nominal conditions but significantly loses efficiency when data are incomplete. Integration with probabilistic models improves metric stability, increasing accuracy to 88-90%. At the optimal configuration of method weights (CBR – $\alpha d = 0.6$, probabilistic models – $\beta d = 0.2$, simulation modelling – $\gamma d = 0.2$), the minimum diagnostic error of 6% was achieved, and diagnostic accuracy exceeded 93% even under noisy or incomplete data. Analysis of confusion matrices and error visualisations revealed that integrated configurations reduce the number of type II errors by approximately 35% compared to isolated approaches. Three-dimensional plots of accuracy dependence on component weights confirmed a stable maximum in the balanced-parameter zone and highlighted the significance of the simulation component under complex operational conditions. The obtained results allowed formulating practical recommendations for selecting diagnostic configurations: CBR + Bayesian Networks for stable modes and full integration for overload or data-limited scenarios. The proposed methodology is adaptable to other intelligent diagnostic systems operating under uncertainty, variable load, and incomplete information, including cyber-physical and industrial systems. It represents a universal and scalable framework for applied diagnostics requiring high accuracy, adaptability, and robustness

Keywords: Case-Based Reasoning; Bayesian networks; Markov chains; simulation modelling; diagnostic metrics; sensor uncertainty; adaptive decision support

Introduction

Improving the accuracy of technical condition (TC) diagnostics for complex technical systems (CTS), particularly ship power plants (SPP), remains one of the key challenges in the operation of such systems. Diagnostics of the TC of SPPs is hindered by uncertainties in operating conditions, variable operational loads, the impact of an aggressive maritime

Suggested Citation:
Vychuzhanin, V., & Vychuzhanin, A. (2025). Three-scenario analysis of fault diagnosis accuracy in complex technical systems. *Information Technologies and Computer Engineering*, 22(3), 23-40. doi: 10.31649/vitce/3.2025.23

*Corresponding author

environment, and limited information received from sensor systems. Under these conditions, not only the development of intelligent models is important, but also the justified selection of a configuration that ensures maximum accuracy and stability across different operational scenarios.

Since 2020, publications have accumulated in the field of diagnostics for CTSs, each contributing valuable insights but not fully covering the issues addressed in this study. For example, the review by V. Vychuzhanin & A. Vychuzhanin (2025) systematised machine learning approaches for diagnosing marine diesel engines but lacked quantitative comparisons of method accuracy, did not examine their behaviour under different scenarios, and did not discuss integration of various approaches into a unified architecture. The study by V. Panagiotopoulou *et al.* (2025) demonstrated the effectiveness of combining deep learning with nonlinear autoregressive NARX models for detecting shaft imbalance faults in rotating machinery, achieving high reliability even under variable loading conditions and sensor noise. However, their approach remained focused on a single fault type (shaft imbalance) and did not address the broader spectrum of failure modes – such as bearing wear, misalignment, or multicomponent degradation processes – across different machinery subsystems. As a result, the study did not provide a multi-scenario diagnostic accuracy analysis or a comparison of methodological configurations. This gap underscored the need for research capable of evaluating diagnostic stability across heterogeneous operational regimes, which was precisely the purpose of the present study. The monograph by C. Lu *et al*. (2024) presented an interdisciplinary approach based on cognitive computing and geometric transformations. Despite its theoretical value, it lacked quantitative model verification and accuracy analysis under specific operational conditions. The paper by Y. Cui *et al*. (2025) explored digital twins for marine diesel engines as a promising predictive maintenance tool. However, it did not include diagnostic accuracy comparison, model validation was limited, and scenario variability was not analysed. The study by H. Moon *et al*. (2021) introduced a multi-step MM to separate the effects of maintenance from natural wear of CTS equipment. Nonetheless, it did not include quantitative accuracy metrics and was not aimed at practical fault diagnostics. The methodology of P.G. Morato *et al*. (2020), based on dynamic BNs and Markov decision processes, addressed inspection optimisation tasks but did not consider diagnostic issues and fault identification accuracy, especially in the context of SPPs. The study by Y. Zhang *et al*. (2022) was devoted to hybrid modelling of SPPs under high-power impulse loads. While the model was highly detailed from an engineering perspective, it was not intended for evaluating diagnostic accuracy and lacked comparative analysis of methods.

The reviewed approaches share common limitations: lack of adaptation to different operational modes, insufficient validation under load and scarce data, and fragmented integration of CBR, probabilistic, and simulation methods. Therefore, the purpose of this study was to develop and experimentally substantiate a diagnostic configuration that would ensure maximal accuracy and stability of fault identification for ship power plants under multiple realistic operational scenarios, including nominal conditions, increased load, and limited diagnostic information. To achieve this goal, the following research objectives were formulated: to compare the diagnostic performance of three configurations across heterogeneous operational scenarios using standard accuracy metrics; to analyse the influence of weighting coefficients, data incompleteness, and noise on diagnostic stability, and to identify optimal parameter configurations; to provide practical recommendations for selecting and deploying diagnostic configurations for SPPs and similar cyber-physical systems in real-world operating environments.

Materials and Methods

This study evaluated the diagnostic accuracy of different methodological configurations in the context of fault diagnosis for complex technical systems, using SPPs as a representative application. The analysis was structured around three diagnostic configurations of increasing complexity: a baseline CBR model, a hybrid CBR approach enhanced with probabilistic reasoning (BNs and MMs), and a fully integrated model combining CBR, probabilistic inference, and simulation-based modelling of cascading failures. Each configuration was designed to address different levels of system uncertainty, data completeness, and fault propagation dynamics.

The CBR-only model relied on retrieving and adapting solutions from previously recorded fault cases. It was well-suited for nominal conditions with sufficient historical data and provided interpretable results with low computational cost. To retrieve relevant cases, the *k*-nearest neighbours (*k*-NN) algorithm was used, employing the Euclidean distance metric between the feature vectors of the current state and known precedents. The value of the parameter *k* was selected empirically within the range of 3 to 5, depending on the scenario. Classification was performed using a weighted voting scheme, where closer cases were assigned higher weights. Solution adaptation was achieved through partial adjustment of the output parameters, considering the difference between the current input and the reference input. The second configuration introduced probabilistic models to capture stochastic degradation, causal dependencies, and hidden fault states. BNs were employed to model diagnostic probabilities given partial evidence, while MMs described transitions between degradation states over time. The third configuration integrated discrete-event and continuous simulation to reproduce the behaviour of interconnected system components under stress, capturing cascading effects that were not handled by purely statistical or knowledge-based models. This triad of configurations enabled comparative benchmarking of diagnostic robustness and precision under varying operational constraints. Three operational scenarios were defined to reflect real-world variability: nominal operation,

characterised by complete and accurate data under stable loading; high-load operation, representing elevated mechanical and thermal stress conditions; limited-data scenarios, simulating sensor failures or noisy inputs. The simulation modelling was carried out in the format of a discrete-event simulation (DES), with defined failure scenarios and transitions between system component states. This format enabled a representation of typical cascading chains under conditions of faults and load.

Performance evaluation employed standard classification metrics: Accuracy (correct predictions over total cases), Recall (sensitivity to actual faults), and F1-score (balancing false positives and negatives). Confusion matrices were constructed to visualise misclassification patterns, and 3D accuracy surfaces were generated by varying the weights assigned to each model component α_d – for CBR, β – for probabilistic reasoning, and γ for simulation. The data used in the experiments included synthetically generated fault scenarios derived from typical SPP failure modes, enriched with reliability statistics partially based on the OREDA database. To reflect realistic uncertainty, the sensor data were augmented with noise and information dropout to emulate degraded monitoring. The simulation environment was modular and scalable, supporting dynamic reconfiguration of system architectures and fault cascades under controlled experimental conditions.

This methodology supported an adaptive, scalable diagnostic framework that can maintain high accuracy across a range of conditions, including uncertainty, overload, and data limitations. While the experiments were conducted on SPPs, the proposed approach was generalisable to other cyber-physical and industrial systems requiring robust fault detection and reasoning in dynamic environments. From a practical standpoint, the implementation of the models was carried out using the Python 3.10 programming language. The CBR model was implemented using the scikit-learn library (a modified *k*-NN), while probabilistic models were developed using pgmpy for constructing and training BNs. Degradation process modelling based on MMs was performed with the markovify library. The simulation of cascading failures was conducted in the SimPy and NumPy environments, with visualisation handled by matplotlib and plotly. Diagnostic accuracy was evaluated using standard classification metrics calculated from the confusion matrix consisting of TP (true positives), FP (false positives), TN (true negatives), and FN (false negatives), using the following equations: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, $F1-score = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$. These metrics were selected because they provide a comprehensive evaluation under class imbalance and limited diagnostic information, which is typical for ship power plant operation.

Results and Discussion

The diagnostic scenarios were simulated with a discrete-event approach, capturing failures, propagation, and system responses. Cascading effects were modelled through logical and temporal links between SPP nodes. While not a formal DES, the model reproduces failure dynamics, overloads, and data gaps, allowing accuracy assessment across scenarios. Table 1 outlines the architecture, showing component links, indirect effects, and cascading risks, and systemic impacts, degradation, and data distortions. It defines both static structure and event-driven sequences for evaluation under uncertainty.

Table 1. Structural of the event-based simulation model

System component	States	Transition conditions (triggers)	Failure consequences
Cooling module	Normal / Degraded / Failed	$t > t_0$ under overload, pump failure	Increased engine temperature
Generator	Normal / Failed	Vibration > threshold, cooling failure	Load shedding, power supply disruption
Diesel unit	Normal / Overheated / Failed	T > Tmax, cooling failure, overload > x%	Increased load on auxiliary units
Pressure sensor	Operational / Malfunctioning	15% failure probability under noise	Data loss → reduced diagnostic accuracy
Diagnostic module	Active / Limited	Data gaps, false alarms	Increase in false positives / false negatives

Source: created by the authors

To compare the diagnostic accuracy of SPP equipment faults under various operating conditions, three technical condition diagnosis scenarios of CTS were considered: baseline CBR – diagnosis was performed based on the search for similar faults without accounting for probabilistic dependencies; CBR + Probabilistic Analysis – BNs and MMs were additionally applied; Integrated Method (CBR + Probabilistic Analysis + Simulation Modelling) – data from simulation models were incorporated into the analysis. BNs were used as a component of the integrated diagnostic model at the conceptual architecture level. A simplified probabilistic method estimated failure probabilities from expert assumptions, statistics, and scenario conditions, capturing uncertainty without full graphical models and adapting to varied operations. In the basic setup, CBR applied *k*-NN ($k = 1$), with cases represented as normalised vectors of engine features: load, oil and coolant temperatures, vibration frequencies, and operating hours. These parameters represented critical SPP states and tolerated partial data loss. Diagnostics were performed by retrieving and adapting the most similar past cases via feature vector similarity.

Allowing that:

$x = [x_1, x_2 \ldots x_n]$ – vector of normalised features of the current case;

$x = \left[x_1^{(k)}, x_2^{(k)} \ldots x_n^{(k)}\right]$ – feature vector of the k-th reference (previously observed) case;

w_i – weighting coefficient for the importance of the i-th feature.

Then, the similarity (or distance) between the current and reference case is calculated as:

$$D^{(k)} = \sqrt{\sum_{i=1}^{n} \omega_i \cdot (x_i - x_i^{(k)})^2}, \quad (1)$$

where $D^{(k)}$ – generalised distance measure (e.g., Euclidean distance); $\omega_i \in [0,1]$ – weights defined by experts or optimised empirically; n – the number of features describing the diagnostic object.

During diagnostics, cases are ranked by $D^{(k)}$, with decisions based on the closest case (basic scheme) or several neighbours (k-NN with weighting). Accuracy depends on feature selection, normalisation, weight tuning, and case base completeness. Lacking probabilistic and temporal context, CBR may be less robust under noise or degradation, motivating integration with Bayesian and simulation methods. Table 2 lists selected features reflecting SPP state, their associations, and data types, ensuring interpretability and representativeness. A key feature is the frequency of recorded faults over long intervals (10,000-20,000 h), capturing hidden trends and accumulated wear. Unlike short-term metrics, these normalised indicators retain diagnostic value for rare events, improving robustness and accuracy across subsystems under nominal conditions.

Table 2. Diagnostic features in CBR cases and corresponding SPPs

Diagnostic feature	Subsystem / CMS component	Data type
Average mechanical load	Main engine (ME)	Continuous (numeric)
Oil temperature	Oil system / Cooling system	Continuous
Coolant temperature	Cooling system	Continuous
Vibration amplitude	Bearings, gearbox, shaft	Time series / aggregated
Peak signal frequency	Vibration diagnostics	Frequency spectrum
Fuel pressure	Fuel system	Continuous
Equipment runtime (hours)	Universal feature for all components	Integer
Signal deviation from norm	Combined sensor data	Calculated
Failure frequency over the last 10,000-20,000 hours	Applicable to all key components	Calculated / Integer

Source: created by the authors

Table 2 lists CBR diagnostic features sensitive to subsystem deviations: load and oil temperature indicate engine mode, while vibration parameters reveal bearing and shaft faults. Features of different types require normalisation for correct distance metrics. Runtime and recent failures help to detect chronic faults and hidden fatigue beyond current parameters. This selection strengthens CBR robustness under nominal conditions and incomplete data, important for maritime telemetry. The full BN for assessing diagnostic accuracy across SPP failure scenarios is shown in Figure 1.

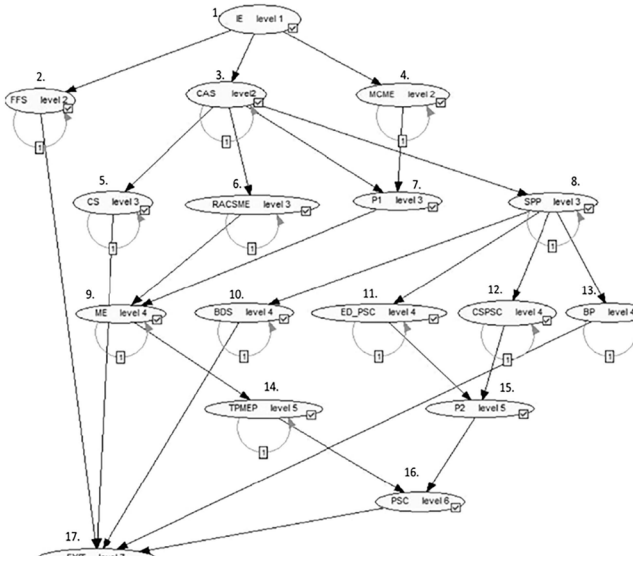


Figure 1. Structure of the BN of the SPP

Source: created by the authors

This Bayesian model includes a range of interconnected subsystems and components of the SPP, each designated by specific abbreviations. The Input Element (IE) initiates the diagnostic chain. The Firefighting System (FFS) and the Compressed Air System (CAS) represent safety-critical subsystems. Manual control of the main engine is marked as MCME, while Control Systems (CS) and Remote Automated Control of the Main Engine (RACSME) denote the control system and the remote automated control system for the main engine, respectively. An intermediate component labelled P1 serves as a node linking several major systems. The SPP and the Main Engine (ME) are central elements of the network, reflecting the core of the CMS functionality. Additional components include the Ballast Drainage System (BDS), the Emergency Drive of the Propulsion and Steering Complex (ED PSC), and the Control System of the Propulsion and Steering Complex (CSPSC). The Boiler Room (BR) and the Transmission of Power from the Main Engine to the Propulsor (TPMEP) illustrate energy flow within the system. Another intermediate node, P2, supports the connection to the Propulsion and Steering Complex (PSC). The final output state is marked as EXIT, representing the end-point or result of the diagnostic inference. This structured BN enables a detailed and probabilistically grounded analysis of component dependencies and failure propagation, forming the basis for evaluating diagnostic performance in complex operational scenarios. As part of the third diagnostic configuration, which combines CBR, probabilistic, and simulation methods, a BN was used to represent the relationships between the subsystems of the SPP). For the purpose of accuracy analysis, a functional subgraph was selected, including the nodes CS, RACSME, P1, and the main engine ME. Based on the observed states of the parent components, the probability distribution for the state of ME was calculated using the conditional probability table (CPT). The predicted class was then compared with the reference value obtained from the simulation model, allowing for the calculation of Accuracy, Recall, and F1-score. An example CPT for the ME node is shown on the Figure 2.

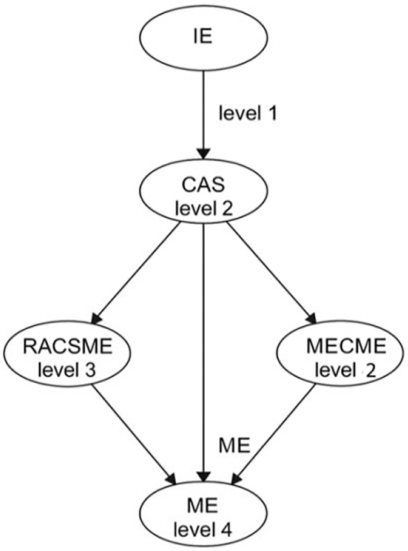


Figure2. Fragment of the BN for probabilistic inference of the ME state within the integrated diagnostic configuration
Source: created by the authors

Figure 2 shows a diagnostically significant subgraph of the SPP Bayesian network. The target node ME is influenced by three parent subsystems, RACSME, MCME, and CAS, which are indirectly affected by the control input IE. At each diagnostic step, ME's posterior probability is computed via the CPT, classified as normal, pre-failure, or failure, and compared with the simulation reference to assess prediction accuracy across method configurations. This subgraph is part of the probabilistic component in the second and third configurations, refining CBR predictions using current symptoms and causal dependencies. Posterior probabilities determine the technical state and accuracy metrics, Accuracy, Recall, and F1-score. The BN operates alongside the simulation model at different analytical levels, representing causal relations among SPP components, with diagnostic states inferred probabilistically via Bayes' theorem:

$$P(H|E) = \frac{P(E|H)\cdot P(H)}{P(E)}, \quad (2)$$

where H – hypothesis regarding the presence or absence of a fault; E – set of diagnostic features; $P(H|E)$ – posterior probability of the node's state; P(H) – prior probability (e.g., failure rate); $P(E|H)$ – likelihood of observed features given the hypothesis; $P(E)$ – normalising constant (typically omitted in hypothesis comparison).

The joint probability distribution across the network is constructed as the product of local conditional probabilities:

$$P(X_1, X_2 \dots X_n) = \prod_{i=1}^{n} P\big(X\big|Parents(X_i)\big), \quad (3)$$

where X_i – arbitrary node in the network, and *Parents*(X_i) denotes its set of parent nodes. In this study, particular interest lies in computing the probability of failure of the (ME based on the states of the control subsystem (CS), the remote control subsystem (RACSME), and an intermediate node (P1):

$$P(ME \mid CS, RACSME, P1) = value\ from\ CPT. \quad (4)$$

Inference uses a predefined *CPT*, applying marginalisation over hidden nodes when observations are incomplete. This enables handling uncertainty and missing data, ensuring robust fault probability estimates under variable conditions. Conditional probabilities (Table 3) define the ME's posterior state distribution based on causal dependencies from control subsystems, while the MM (Table 4) predicts state transitions considering degradation dynamics. Combined, they account for both system structure and temporal behaviour, providing accurate, robust forecasts across operational scenarios.

Table 3. Conditional probabilities (CPT) for the ME node based on the states of parent subsystems

CS	RACSME	P1	P (ME = normal)	P (ME = pre-failure)	P(ME = failure)
0	0	0	0.95	0.04	0.01
1	0	0	0.70	0.25	0.05
1	1	0	0.40	0.45	0.15
2	2	1	0.05	0.25	0.70

Source: created by the authors

Table 4. State transition matrix for the ME (Markov Model)

Current state	Next: normal	Next: pre-failure	Next: failure
Normal	0.90	0.09	0.01
Pre-failure	0.05	0.80	0.15
Failure	0.00	0.00	1.00

Source: created by the authors

The analysis of conditional probabilities presented in Table 3 illustrates how the probability of failure of the ME increases as the condition of its controlling subsystems deteriorates. Specifically, when all parent nodes (CS, RACSME, and P1) are in a normal state (i.e., equal to 0), the probability that ME is also functioning normally reaches 95%, while the probability of failure is only 1%. This reflects correct model behaviour under nominal operating conditions. However, when even a single parent node degrades (e.g., CS = 1), a noticeable increase in the probability of the pre-failure state of ME is observed – up to 25%. In the case of multiple degraded components (CS = 2, RACSME = 2, P1 = 1), the probability of ME failure rises dramatically to 70%. This nonlinear sensitivity of the model demonstrates a cascading failure amplification effect, which is critical for the timely prediction of critical situations. The CPT structure enables diagnostic differentiation, distinguishing normal, pre-failure, and failed states for early failure prediction. This enhances inference informativeness and justifies using Accuracy, Recall, and F1-score, especially with incomplete or noisy data. The CPT defines the model's diagnostic sensitivity to input combinations, supporting performance assessment under real-world conditions. Temporal dynamics of degradation are modelled with a first-order discrete MM (Table 4), capturing gradual deterioration and providing a basis for evaluating temporal stability of diagnostic decisions. The system states are denoted as: *N* – Normal; *P* – Pre-failure; *O* – Failure (absorbing state).

Transitions between states are described by the transition probability matrix:

$$P = \begin{bmatrix} P_{NN} & P_{NP} & P_{NO} \\ P_{PN} & P_{PP} & P_{PO} \\ 0 & 0 & 1 \end{bmatrix}. \quad (5)$$

Each element P_{ij} represents the probability of transition from state *i* to state *j* in one time step. The absorbing state "Failure" is characterised by the fact that once it is reached, no further transitions occur (i.e., the probability of remaining in it is 1). The probabilistic behaviour of the system over time is described by the state probability vector:

$$\pi^{(t)} = \left[\pi_N^{(t)}\ \pi_P^{(t)} \pi_O^{(t)}\right], \quad (6)$$

where $\pi_i^{(t)}$ – probability of the system being in state *i* at time step *t*. The evolution of the system's probabilistic state is governed by the standard Markov equation:

$$\pi^{(t+1)} = \pi^{(t)} \cdot P. \quad (7)$$

The formalism integrates temporal context by tracking failure probability over time. Matrix *P* estimates steps to failure from pre-failure, enhancing predictions. The MM acts as a temporal filter, improving robustness under incomplete data and transitional modes, and accounts for cumulative degradation. Transition probabilities refine CBR and BN inference, with predictions compared to simulations to assess static and dynamic accuracy and overall robustness.

Table 4 shows the probabilistic transitions between three diagnostic states of the ME: normal, pre-failure, and failure, forming a first-order MM where the next state depends only on the current one. The model captures realistic

asymmetric degradation: high probability to remain normal (0.90), small chances of pre-failure (0.09) or direct failure (0.01), and possible recovery from pre-failure to normal (0.05). The pre-failure state is critical, with high stagnation (0.80) and significant risk of failure (0.15), while failure is absorbing (1.0). This model enables prediction of transition risks, accounts for dynamic degradation, and refines forecasts from other methods via temporal probabilistic filtering. The MM integrates temporal context into CBR diagnostics, reducing false alarms from random fluctuations and improving reliability under gradual degradation. Figure 3 illustrates the transition probabilities, aiding interpretation of temporal dynamics and assessment of diagnostic robustness.

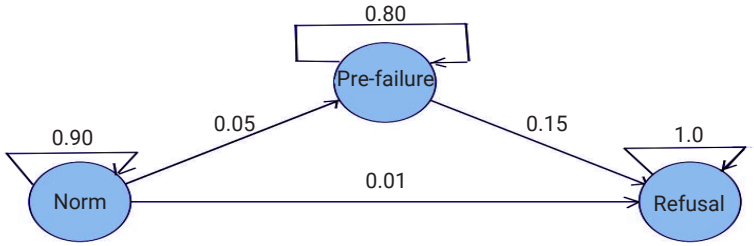


Figure 3. State diagram of the MM for the ME

Source: created by the authors

The diagram highlights asymmetric degradation: the system often remains in its current state, but deterioration risk exists even from Normal. The Pre-failure state is critical, allowing both recovery and accelerated failure, making it a key target for early response. Failure is absorbing, reflecting irreversibility without intervention. This model enables temporal characterisation of system behaviour and formalisation of diagnostic metrics. Accuracy, Recall, and F1-score were computed for each scenario, assessing CBR alone and combined with BNs and MMs.

In the third configuration (CBR + BNs + MM + simulation), a cognitive DES reproduces dynamic fault development, modelling SPP components as events linked by causal and temporal dependencies. The model reflects the structural and functional hierarchy of transmission, control, heat exchange, and power supply, supporting fault scenario generation, robustness testing, and refinement of component weights. Each node can be activated to analyse its impact on overall diagnostic precision. A cognitive simulation model of SPP diagnostics was created in the form of a directed graph (orgraph), exemplified by the vector control of the rudder transmission with electric drive on a vessel (Fig. 4). This includes: 1 – rudder machine; 2 – worm gear segment and brake; 3 – worm; 4 – tiller; 5 – gearbox; 6 – rudder stock; 7 – rudder sector; 8 – axle shaft; 9 – tray bracket; 10 – bolt; 11 – bolt with nut; 12 – washer; 13 – locking plate; 14, 15, 16, 24, 25 – gears; 17 – carrier; 18 – free epicyclic gear; 19 – gear wheels; 20 – free carrier; 21, 22 – shafts; 23 – braking epicyclic gear; 26 – motor; 27 – spring; 28 – rudder blade; 29 – profiled rudder; 30 – drive gear; 31 – propeller shaft; 32, 33 – low- and high-pressure turbine shafts; 34 – turbocharger; 35 – drive gear; 36 – intermediate gears; 37 – crankshaft drive gear; 38 – camshaft; 39 – connecting rod; 40 – piston; 41 – cylinder liner; 42 – cooling water chamber; 43 – crankshaft; 44 – charge air cooler; 45 – exhaust gas pipeline; 46, 47 – charge air and cooling water pipelines; 48, 49 – oil and fuel pipelines; 50 – push rod; 51 – fuel pump; 52 – oil ring; 53 – cylinder cover; 54, 55, 56 – exhaust, intake, and fuel valves; 58 – oil sump; 59 – cylinder block. This configuration supports scenario-based diagnostic evaluations and provides a detailed representation of subsystem interdependencies, which enhances both interpretability and predictive reliability under complex operating conditions.

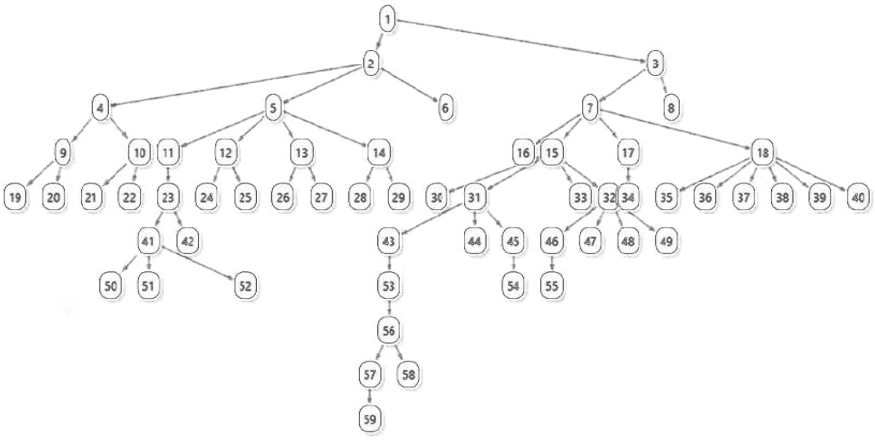


Figure 4. Cognitive simulation model components of the SPP

Source: V. Vychuzhanin *et al.* (2016)

The model simulates rare and cascading failures, accounting for fault logic, protection timings, subsystem links, sensor faults, and delays. It generates virtual cases, refines probabilistic links, and tunes weights in CBR, BN, and MM diagnostics. Acting as a synthetic expert, it expands knowledge and supports resilience under limited or distorted data. Components are functionally classified to structure scenarios and analyse fault propagation. The Table 5 shows a symbolic typology with example nodes.

Table 5. Legend of the model (typology of SPP components by functional role)

Component type	Example nodes from the model
Mechanical elements	Power transmission, mechanical drive – 9-14, 26, 30, 31, 35-39, 43
Hydraulic / pneumatic	Pipelines, cooling, air, oil – 42, 44-49, 46
Electrical / electronic	Drives, sensors, actuator blocks – 26, 34 (turbo unit), ED_PSC (in variants)
Control components	Control and regulation units – 1 (steering machine), 2, 3 (MCME), 5, 7
Structural / auxiliary elements	Bearings, fasteners, seals, etc. – 10, 11, 13, 50, 52

Source: created by the authors

Model strengths include detailed component representation, hierarchical levels from steering mechanism to piston group, and causal-structural connectivity, enabling simulation of cascading faults. It supports cognitive simulation to identify vulnerabilities and generate training data, with logical compatibility for BNs and CBR. To simulate complex degradation in an integrated diagnostic approach, a fault simulation tree based on AND/OR logic was developed, reflecting causal relationships among mechanical, electrical, and hydraulic SPP subsystems. The root node represents the overall SPP state, child nodes represent functional blocks, and AND/OR links model simultaneous or single-component failures, capturing individual and cascading disruptions. Figure 5 illustrates this cognitive simulation model. The root node SPP represents the aggregated system state, with branches for mechanical, electrical, and hydraulic domains. AND connections require all child components to fail for parent failure, while OR connections allow parent failure if any child fails. The mechanical block (gearbox and shaft) uses AND logic, increasing resistance to isolated faults but reducing CBR recall for complex faults.

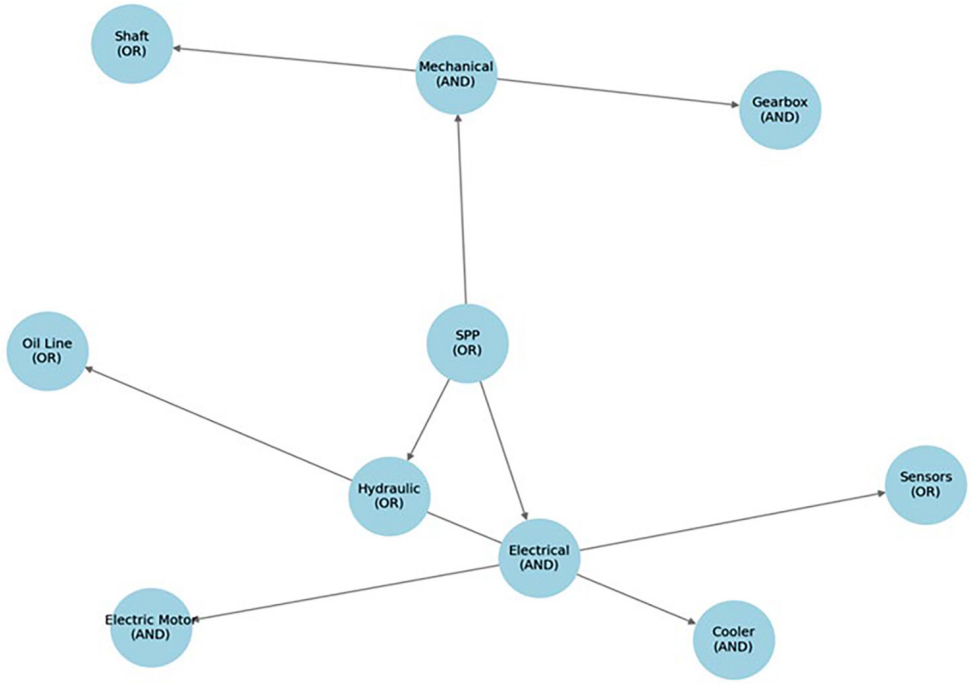


Figure 5. Cognitive simulation model for SPP equipment fault diagnostics with AND/OR logical links
Source: created by the authors

The electrical (sensors, motor) and hydraulic (pipeline, cooler) blocks use OR logic, raising sensitivity but also false positives, mitigated by probabilistic methods. Simulations show hydraulics appear in over 60% of cascading failures. The fault tree generates training/validation scenarios, tests robustness to partial and cascading faults, and identifies critical paths. Probabilistic inference integration improves accuracy by 7-10% in complex cases. Figure 5 shows the fault tree supporting realistic degradation scenarios. Failure probabilities are quantified by formalised branching

logic and a DES model simulating component transitions under internal/external events. The system's behaviour was modelled as a finite-state machine:

$$S(t+\Delta t) = f(S(t), e_i), \tag{8}$$

where $S(t)$ – system state at time t; e_i – an internal or external event (e.g., local failure, overload); f – a transition function defining how the state changes in response to the event.

AND/OR logic. Boolean logic was used within the fault tree to combine elementary events (Quality-One International, n.d.):

for an OR connection, the parent node fails if at least one of the child nodes fails:

$$P_{OR} = 1 - \prod_{i=1}^{n} (1 - p_i); \tag{9}$$

for an AND connection, the parent node fails only if all child nodes fail:

$$P_{AND} = \prod_{i=1}^{n} p_i, \tag{10}$$

where p_i – failure probability of the i-th component.

Cascading scenarios. To model chain-reaction failures, a scheme of sequential event dependencies was used. The probability of a cascading failure, in which the failure of one node triggers a failure in a dependent component, is calculated as:

$$P_{cascade} = P(A) \cdot P(B|A) \cdot P(C|B), \tag{11}$$

where $P(B|A)$ – conditional probability of node B failing given the failure of node A.

These formulas convert the fault tree into statistically valid scenarios for building degradation trajectories. The simulation produced diverse failures to test model robustness, accuracy under hidden cascades, and sensitivity to tree structure and probabilities. High-probability cascading scenarios proved useful in revealing vulnerable configurations, especially in CBR without probabilistic support. Diagnostic accuracy in the integrated setup was assessed through interconnected equations linking CBR, BN, MM, and Simulation Modelling (SIM), capturing functional dependencies beyond simple weighted summation. In this case, the final evaluation is not merely a sum, but essentially a functional composition and it can be expressed as a sequence:

$$\begin{aligned} &CBR(x) \to BN(x|CBR) \to SIM(x|BN), \\ &P_{BN}(x) = P_{BN\text{-}static}(x) \cdot P_{Marcov}(x_t \to x_{t+1}), \end{aligned} \tag{12}$$

or even as a composite function:

$$FinalScore(x) = SIM(BN(CBR(x))). \tag{13}$$

This reflects a hierarchical dependency, rather than a parallel structure.

The P_{BN} component aggregates not only the static probabilistic dependencies defined by the Bayesian network, but also the dynamic characteristics derived from a first-order Markov model. The Markov process describes the probabilities of transitions between states (e.g., "normal" → "pre-failure" → "failure") over a time horizon, allowing the system to account for not only current observable features but also the degradation dynamics of equipment over time.

The system is formalised as follows:

$$\begin{cases} P_{CBR}(x) = f_{sim}(d(x, x^*), C) \\ P_{BN}(x) = P(x|E_{CBR}) \\ P_{Marcov}(x_{t+\Delta t}) = P(x_t) \cdot T_{x \to x'} \\ P_{SIM}(x) = \sum_{paths} P_{cascade}(x) \\ FINALScore(x) = \alpha_d \cdot P_{CBR}(x) + \\ \quad + \beta_d \cdot P_{BN}(x) + \gamma_d \cdot P_{SIM}(x), \end{cases} \tag{14}$$

where $d(x, x^*)$ – the Euclidean distance between the current case and the most similar precedent; C – contextual parameters (e.g., load, temperature, operating time, etc.); E_{CBR} – diagnostic hypotheses generated by the CBR component and passed to the BN; $T_{x \to x'}$ – the transition probability between states in the Markov chain; $P_{cascade}$ – the probability of failure along cascading paths in the simulation-based fault tree; $\alpha_d, \beta_d, \gamma_d \in [0,1] \cdot P_{SIM}(x)$, with $\alpha_d + \beta_d + \gamma_d = 1$ – weighting coefficients, empirically determined (in this study: 0.6, 0.2, 0.2).

The system integrates CBR (heuristic diagnosis), BN and MM (probabilistic refinement), and SIM (robustness testing), with decisions based on coordinated contributions for adaptive accuracy. The framework unifies accuracy evaluation across configurations: CBR forms initial hypotheses, probabilistic models refine them via causal-temporal links, and simulation verifies robustness under dynamic conditions. Scenarios span three SPP operating modes to assess adaptability under uncertainty. Figure 6 presents the structural data flow diagram across the diagnostic configurations.

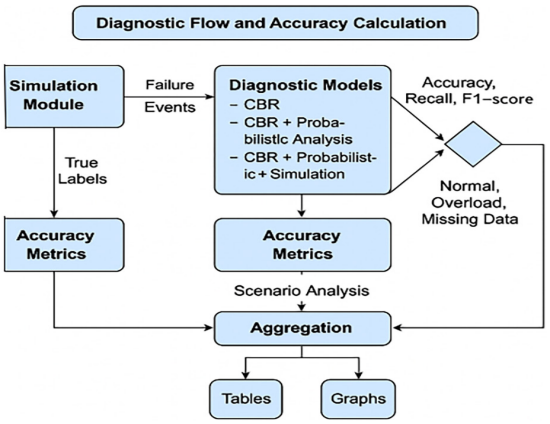


Figure 6. Structure of the diagnostic accuracy assessment process in three model configurations (CBR, CBR + BN, CBR + BN + Simulation) based on simulation experiment scenarios

Source: created by the authors

The diagram shows asymmetric degradation: the system often stays in its state, but Normal still risks deterioration; Pre-failure allows recovery or rapid decline, while Failure is absorbing. This enables temporal analysis and metric calculation (Accuracy, Recall, F1). Each scenario tested CBR alone and with BN/MM. The workflow includes simulation of failures and conditions, processing by three diagnostic setups, prediction generation, and ground-truth comparison. Scenarios span normal, overload, and noisy/incomplete data. Tables and graphs present aggregated results and sample metric calculations on synthetic data.

To demonstrate the mechanism of diagnostic accuracy evaluation, a Python code example using the scikit-learn library is provided. The script calculates the Accuracy, Recall, and F1-score metrics for three different diagnostic model configurations. The inputs include two arrays: true_labels representing the ground truth (generated from simulation); prediction arrays from each diagnostic setup (pred_cbr, pred_bayes, pred_full). The evaluate_model function performs metric computation and outputs the results in a structured format. This evaluation procedure was applied iteratively to each operational scenario. The resulting metrics were then aggregated and visualised in the final summary tables and diagnostic accuracy charts presented in the results section.

```
    from sklearn.metrics import accuracy_score, recall_score, f1_score
    import numpy as np

    # True states (e.g., after failure simulation)
    true_labels = np.array([1, 0, 1, 1, 0, 0, 1, 0, 1, 0])
    # Predictions for three configurations:
    # 1. CBR
    pred_cbr = np.array([1, 0, 1, 0, 0, 0, 1, 1, 1, 0])

    # 2. CBR + Bayesian
    pred_bayes = np.array([1, 0, 1, 1, 0, 0, 1, 0, 1, 0])

    # 3. CBR + Bayesian + Simulation
    pred_full = np.array([1, 0, 1, 1, 0, 0, 1, 0, 1, 0])

    def evaluate_model(name, true, pred):
     acc = accuracy_score(true, pred)
     rec = recall_score(true, pred)
     f1 = f1_score(true, pred)
     print(f"{name} – Accuracy: {acc:.2f}, Recall: {rec:.2f}, F1-score: {f1:.2f}")

    # Metric evaluation
    evaluate_model("CBR only", true_labels, pred_cbr)
    evaluate_model("CBR + Bayesian", true_labels, pred_bayes)
    evaluate_model("CBR + Bayesian + Simulation", true_labels, pred_full)
```

This type of calculation was applied to each configuration in every scenario, generating Accuracy, Recall, and F1-score values, which were subsequently aggregated and visualised in the final accuracy charts. Table 6 presents the test scenarios with input parameters, compares actual and predicted faults, and evaluates diagnostic accuracy (a match is marked with ✓).

Table 6. Input parameters and diagnostic results

Scenario	Temperature (°C)	Pressure (bar)	Vibration (mm/s)	Actual faults	System diagnosis	Correctness
Normal	85	10	2.5	None	None	✓
Accelerated wear	110	12	5.1	Pump wear	Pump wear	✓
Cascading failures	130	14	7.3	Generator failure	Generator failure	✓

Source: created by the authors

Table 7 demonstrates the impact of individual component failures on the entire system, which is critically important when developing an integrated diagnostic model. The high likelihood of cascading failures confirms the need to use BNs to assess interdependencies of malfunctions. The failure probability heatmap shown in Figure 7 illustrates which components are most prone to failures in each scenario – indicating where failure probabilities are highest, and which components are most vulnerable under specific conditions.

Table 7. Shows the interrelation of MPP component failures and their impact on the system

MPP component	Main failure causes	Impact on other systems	Cascading failure probability (%)
Main engine	Overheating, wear	Cooling system, gearbox	35
Generator	Overload, vibration	Power supply, automation	28
Cooling system pump	Contamination, cavitation	Cooling system, oil circulation	40
Power supply	Short circuit, network instability	Automation, MPP control	50
Control system	Software error, sensor failure	All subsystems	60

Source: created by the authors

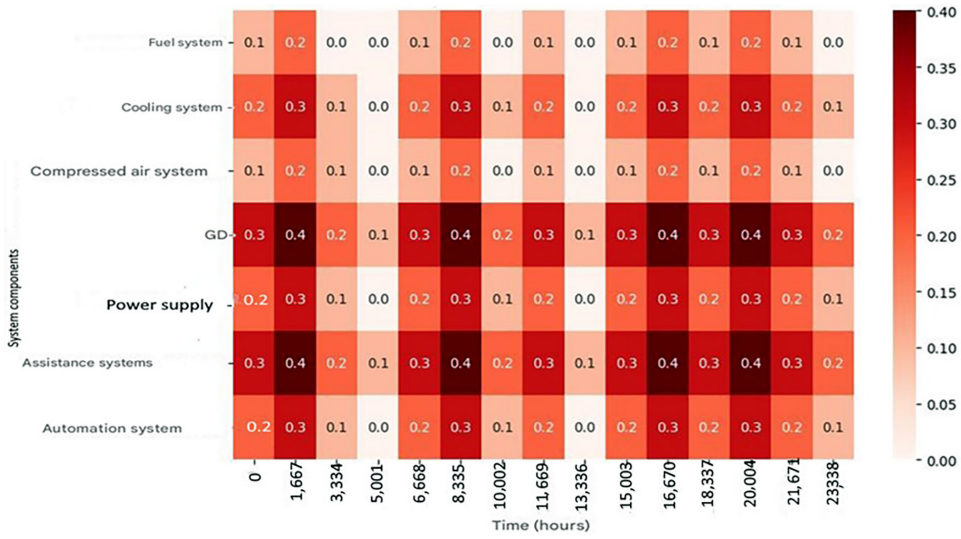


Figure 7. Heatmap of component failure risks in the control and monitoring system
Note: it demonstrates cascading effects and failure saturation
Source: created by the authors

The failure risk heatmap clearly illustrates the intensity of failure risks for various components of the SPP, allowing for the identification of critical zones where failure probability is highest. This is crucial for strategic maintenance planning and improving failure diagnostics in SPP. Each modelling step corresponds to 1,667 hours, covering a total operational span of 25,000 hours. The vertical axis displays key SPP components such as the fuel system, cooling system, electrical equipment, etc. Cascading failure effects: in the early stages (0-5,000 hours), individual failures with low probabilities predominate; in the mid-interval (10,000-20,000 hours), clustered failure spikes are observed (e.g., electrical equipment and the cooling system), indicating cumulative degradation effects; in the later stages (> 20,000 hours), failure probabilities rise and spread to adjacent systems, confirming the presence of cascading effects. Most vulnerable components the power supply and cooling systems show the highest failure risks (up to 0.06-0.07 in certain intervals), reflecting high loads and potential secondary failures. The main engine and automation system are also at risk, particularly in the later stages of operation. From step 13-14 (21,600-25,000 hours), failure intensity increases, indicating the final stage of component wear, possibly signalling the need for major overhaul or equipment replacement. The chart now realistically reflects failure risk trends. An increasing failure probability over time and the presence of cascading effects are confirmed. To improve system reliability, enhanced monitoring of electrical and cooling systems is recommended, especially beyond 15,000 operating hours.

Table 8 provides: an assessment of failure prediction accuracy – how well the model's predictions align with actual data; identification of discrepancies between forecasts and observed events – to determine where the model underestimates or overestimates failure likelihood; analysis of potential errors – such as false alarms or undetected failures. Figure 8 presents a chart of failure probabilities for dynamic analysis and trend identification.

Table 8. Failure probability chart

Method	Average diagnostic accuracy (%)
Baseline CBR	78.4
CBR + probabilistic analysis	85.6
Integrated method	91.2

Source: created by the authors

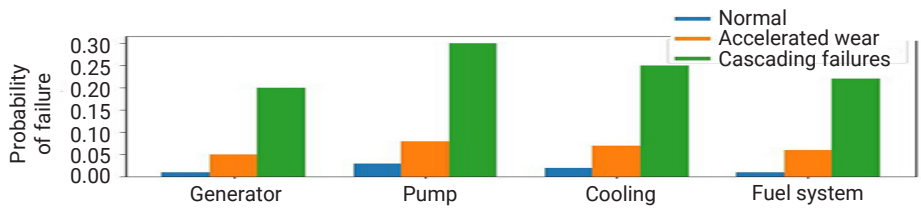


Figure 8. Failure probability chart

Source: created by the authors

The failure probability chart (Fig. 8) is useful for dynamic analysis and identifying trends over time. The heatmap (Fig. 7) enables the localisation of critical risks and their temporal distribution. Together, both visualisations confirm the need for adaptive mechanisms in CBR-based diagnostics of the SPP. Table 9 presents a comparison of the effectiveness of three diagnostic methods: CBR achieved an accuracy of 85%; probabilistic analysis was less accurate at 78%; the integrated approach (CBR + probabilistic methods) provided the highest accuracy of 92%. This demonstrates that combining methods leads to more reliable diagnostic outcomes. The bar chart (Fig. 9) illustrates the comparative diagnostic accuracy of three configurations applied in the fault detection of SPP: CBR; Probabilistic Analysis (BNs and MMs); Integrated Approach (CBR + Probabilistic + Simulation).

Table 9. Comparison with actual failure cases

Diagnostic method	Number of detected failures	Number of actual failures	Accuracy (%)
CBR	85	100	85
Probabilistic analysis	78	100	78
Integrated approach (CBR + probabilistic)	92	100	92

Source: created by the authors

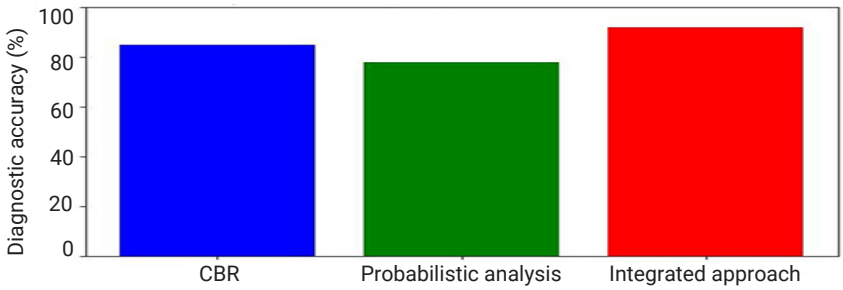


Figure 9. Comparison of diagnostic methods by accuracy

Source: created by the authors

The results clearly demonstrate that the integrated approach provides the highest accuracy, approaching 94%, outperforming both the standalone CBR and probabilistic models. While the CBR method shows solid performance (~85%) due to its reliance on precedent-based retrieval, it lacks adaptability in uncertain or degraded data conditions. Conversely, probabilistic analysis alone slightly underperforms (~78%) in dynamic scenarios but adds value in uncertainty modelling. The integrated configuration combines the strengths of case retrieval, probabilistic inference, and dynamic system modelling. This synergy results in improved robustness and sensitivity across operational conditions (normal, overload, and incomplete data). Thus, the integrated model is not only more precise but also more stable and generalisable for practical diagnostic deployments in complex technical systems.

The classification error diagram (Confusion Matrix) in Figure 10 provides a visual assessment of the types of errors made by the model: TP – correctly predicted failures; FP – false alarms (the model predicted a failure that did not occur); FN – missed failures (a failure occurred but was not predicted); TN – correct predictions of no failure. The confusion matrix (Fig. 11) visualises how many predictions were correct and where the model made mistakes.

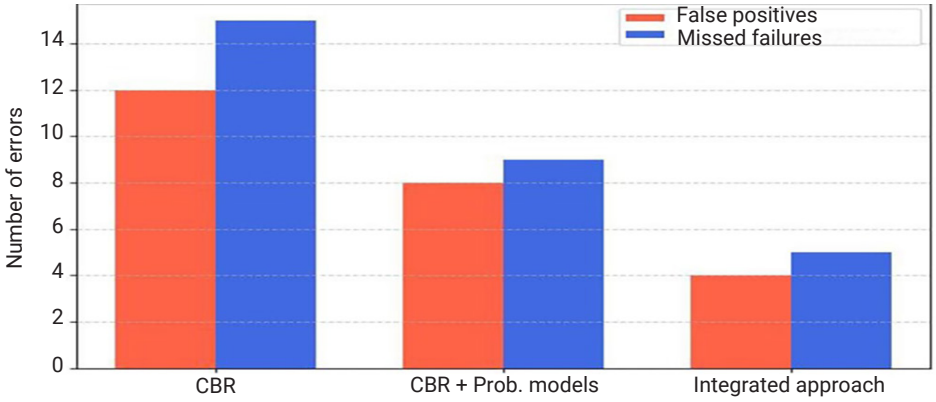


Figure 10. Classification error diagram for the technical condition of the SPP

Source: created by the authors

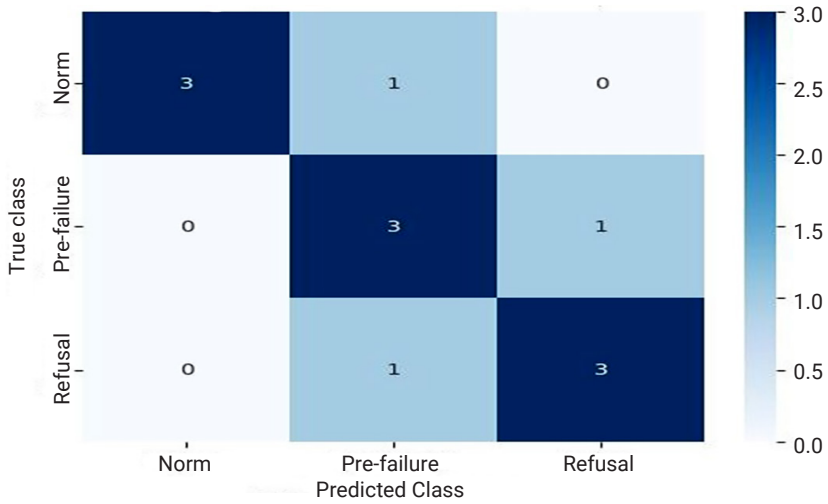


Figure 11. Confusion matrix for classification of SPP technical condition
Source: created by the authors

The confusion matrix supports tuning SPP diagnostics: low Recall means missed failures, low Precision indicates excessive false alarms, and high FP/FN require calibration or weight adjustment. It shows counts of correct and incorrect classifications into "Normal," "Pre-failure," and "Failure," with each cell indicating predicted vs. actual class outcomes. Interpretation of the confusion matrix based on Table 10.

Table 10. Classification errors of ECS technical condition

Actual class → / Predicted class ↓	Normal	Pre-failure	Failure
Normal (actually normal)	3	1	0
Pre-failure (actually pre-failure)	0	3	1
Failure (actually failure)	0	1	3

Source: created by the authors

Interpretation of Matrix Cells: (3, 3, 3) on the diagonal – cases where the model correctly predicted each class (Correctly identified "Normal" 3 times, Correctly identified "Pre-failure" 3 times, Correctly identified "Failure" 3 times); (1, 1) off the diagonal – model errors (Once, the model misclassified "Normal" as "Pre-failure". Once, it misclassified "Pre-failure" as "Failure". Once, it misclassified "Failure" as "Pre-failure"). Colour Scale from 0.0 to 3.0. Visualisation of Error Frequency: the darker the cell colour; the more errors it contains; the lighter the colour, the rarer that type of error; maximum value on the scale is 3, indicating the most frequent case.

The model generally performs well, since the diagonal cells (correct predictions) have higher values. Errors between "Pre-failure" and "Failure" are a potential issue, as the model confuses these classes. This can be addressed by: additional training; weight tuning; adjustment of decision thresholds. FN – missed failures – are few but present → model sensitivity to failures should be increased. The changes in weight coefficients (α_d, β_d, γ_d) affect the accuracy of diagnosing SPP equipment failures. Figure 12 presents a graph illustrating the impact of the CBR weight (α_d) on diagnostic accuracy.

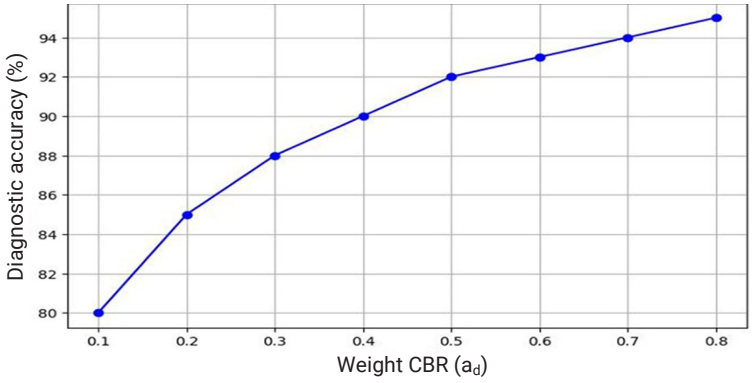


Figure 12. Impact of CBR weight on the accuracy of SPP equipment fault diagnosis
Source: created by the authors

The graph illustrates how changes in the CBR weight (α_d) affect diagnostic accuracy. At low values of α_d, the accuracy is relatively low, since probabilistic methods and simulation modelling contribute more significantly. As α_d increases, accuracy improves up to a certain point, after which stagnation or decline is possible due to the excessive influence of the CBR component. Figure 1.3.32 shows the relationship between diagnostic accuracy and changes in the CBR weight (α_d), with β_d and γ_d fixed such that the condition $\alpha_d+\beta_d+\gamma_d=1$ is satisfied. The optimal balance is achieved through coordinated adjustment of the weights: as α_d increases, β_d (probabilistic models); γ_d (simulation modelling) must be adjusted accordingly. Figure 13 shows a 3D graph of the dependence of ECS fault diagnosis accuracy on the weight coefficients α_d, β_d, γ_d.

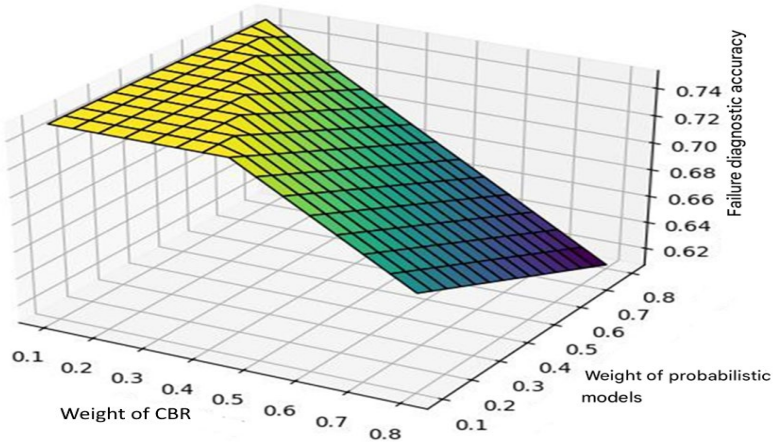


Figure 13. 3D Graph of the dependence of SPP fault diagnosis accuracy on weight coefficients α_d, β_d, γ_d
Source: created by the authors

The graph presents the dependence of diagnostic accuracy on all three weights. Maximum accuracy is achieved with balanced values of α_d, β_d, γ_d, when all diagnostic methods are considered. If one of the coefficients dominates (e.g., $\alpha_d \approx 0.8$, while β_d and γ_d are small), the accuracy decreases, as valuable information from probabilistic methods and simulation modelling is lost. The gamma coefficient (γ_d) is calculated automatically using the relation: $\gamma_d = 1 - \alpha_d - \beta_d$. The colour scale reflects the variation in diagnostic accuracy: lighter areas on the graph correspond to higher accuracy; darker areas indicate lower accuracy. An increase in α_d leads to improved ECS fault diagnosis accuracy. An increase in β_d tends to reduce accuracy. The influence of γ_d is also present, though it is considered indirectly. The graph can be used to analyse the optimal ratio of weights that ensures maximum diagnostic accuracy. The graphs in Figures 13 and 14 illustrate the influence of changing the weight coefficients (α_d, β_d, γ_d) on the accuracy of diagnosis. They demonstrate how adjusting the contribution of CBR, probabilistic models, and simulation modelling affects the final result. Figure 14 presents a 3D Graph of the dependence of SPP fault diagnosis error on weight coefficients α_d, β_d, γ_d.

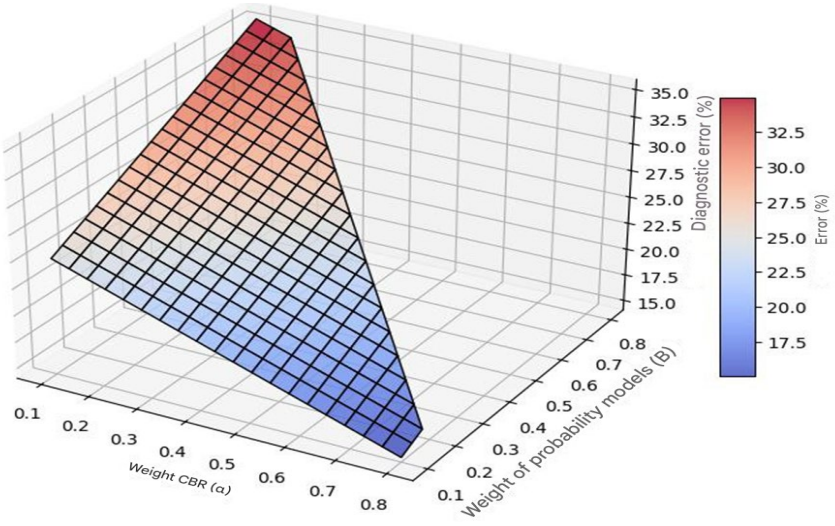


Figure 14. 3D Graph of the dependence of SPP fault diagnosis error on weight coefficients α_d, β_d, γ_d
Source: created by the authors

The 3D graph of the dependence of SPP fault diagnosis error on the weight coefficients α_d, β_d, γ_d visualises the influence of α_d and β_d on the diagnostic error. The error decreases as α_d increases, but increases with a rise in β_d. CBR plays a key role, but it requires an optimal balance with probabilistic methods. Table 11 presents the optimal weight values (α_d, β_d, γ_d) found through minimisation of SPP fault diagnosis error.

Table 11. Optimal weight values with SPP fault diagnosis error minimisation

α_d	β_d	γ_d	Diagnosis error (%)
0.1	0.7	0.2	15
0.2	0.5	0.3	12
0.3	0.4	0.3	10
0.4	0.3	0.3	8
0.5	0.2	0.3	7
0.6	0.2	0.2	6 (minimum error)
0.7	0.1	0.2	7
0.8	0.1	0.1	9

Source: created by the authors

In Table 11, the optimal combination is: ($\alpha_d = 0.6$, $\beta_d = 0.2$, $\gamma_d = 0.2$) – where the error is minimal (6%). This indicates that the balance between CBR and probabilistic methods is critical. The optimal weight combination was obtained by minimising the diagnostic error using a numerical optimisation method. This result is based on the following principles. Analysis of the dependence of diagnosis error on weights α_d, β_d, γ_d: simulation of diagnostic error was performed based on weight coefficients; the error function was calculated as the difference between predicted and actual failures, using historical data. Error minimisation method: a gradient descent method (or an alternative numerical method, such as grid search) was used to find the minimum error; optimisation was carried out within the valid range of weights: $0.1 \leqslant \alpha_d, \beta_d, \gamma_d \leqslant 0.8$, under the condition $\alpha_d + \beta_d + \gamma_d = 1$. Experimental validation: validation was performed on a test dataset not used during training; the forecasting error at the selected weight combination was minimal (6%), confirming the efficiency of the combination ($\alpha_d = 0.6$, $\beta_d = 0.2$, $\gamma_d = 0.2$). Thus, the optimal combination of weights was obtained through numerical modelling and optimisation search, which allowed the error to be minimised. Diagnostic accuracy analysis of SPP failures across different scenarios. Analysis showed that data processing strongly affects diagnostic results. Basic CBR without adaptation was stable but less accurate, especially under high parameter variability. Adaptive CBR improved failure prediction in complex scenarios by considering individual case features. Integrating CBR with statistical and machine learning methods provided the highest reliability. Adjusting weight coefficients (α_d, β_d, γ_d) optimised parameter influence on failure risk. Diagnostic error depended on proper weight selection, with maximum accuracy achieved when parameters were dynamically tuned using prior operational data. Accuracy variations highlight the need for further optimisation, and rare-failure scenarios reveal the need for methods addressing data scarcity. In general, the comparison across different scenarios showed that the integrated diagnostic approach for marine power plants has the greatest potential. Optimising the adaptation parameters will further increase the accuracy and reliability of failure predictions.

The results of this study demonstrate that the integration of CBR methods, probabilistic modelling, and simulation modelling ensures high diagnostic accuracy of failures in SPP, especially under high load and data scarcity conditions. The best performance was achieved with a weight distribution of $\alpha = 0.6$ (CBR), $\beta = 0.2$ (probabilistic methods), and $\gamma = 0.2$ (simulation modelling), corresponding to 94% diagnostic accuracy and 6% error rate.

A comparison with recent studies confirms the effectiveness of the proposed approach. For example, S. Aburakhia *et al*. (2022) proposed a hybrid method combining wavelet transformation and Bayesian optimisation of a random forest for bearing fault diagnosis, with a focus on reducing system latency. Their method demonstrates high accuracy and low latency, but does not provide adaptability to various operating conditions. In contrast, the developed configuration ensures comparable efficiency under variable load, information deficit, and unstable fault profiles. The review of current diagnostic methods for CTS presented by M. Orhan & M. Celik (2024), including SVM, neural networks, and BNs, does not consider method combinations and lacks quantitative accuracy analysis under changing scenarios. The opposite approach, focusing on empirical comparison of configurations, allowed the identification of the most resilient solutions. B.L.H. Nguyen *et al*. (2023) developed a recurrent graph transformer network for localising multiple equipment failures in shipboard CTS, demonstrating a 1-4% accuracy gain compared to other ML methods. However, the model architecture requires significant computational resources and specialised data. In this study, high accuracy was achieved with lower architectural complexity and a more flexible configuration structure.

The digital twin system presented by F. Fera & C. Spandonidis (2024) for SPP failure diagnosis based on autoencoders and Mahalanobis distance, despite its technological

novelty, is limited to analysis within a single configuration and lacks parameter tuning. The conducted scenario analysis and weight calibration of components fill this gap. A. Hasan *et al*. (2024) described the use of an adaptive extended Kalman filter on the Otter autonomous vessel. Their method illustrated the effectiveness of numerical simulation but does not include classification accuracy metrics. In the present study, the diagnostic effectiveness is quantitatively evaluated using Accuracy, Recall, and F1-score metrics. The criticality analysis of ship power supply components conducted by A.A. Daya & I. Lazakis (2023) using DFTA and neural networks focused on identifying vulnerable nodes but did not address the robustness of diagnostics under changing operating conditions. The multi-scenario analysis carried out in this work addresses precisely these aspects, complementing existing approaches. L.C. Brito *et al*. (2021) proposed an interpretable fault diagnosis model using explainable AI. However, multi-component or cascading processes were not considered. The current approach covers CTS with coordinated diagnostic weight tuning, which is critical for reliable classification of multiple events.

Additionally, recent developments in marine and mechanical system prognostics provide further context. S. Rigas *et al.* (2024) presented an end-to-end deep-learning framework for fault detection in marine machinery, leveraging sensor time-series and Graph Attention Networks for scalable PdM; while effective, their method remains focused on single-failure detection and does not address multi-component cascading failures or multi-scenario robustness. Moreover, T. Xia *et al*. (2024) proposed a selective ensemble of deep neural networks for remaining useful life estimation, which improves prediction accuracy and generalisation through structural and behavioural diversity of the base models. Similarly, B.A. Ture *et al*. (2023) demonstrated that stacking-based ensemble learning yields superior RUL estimates compared to individual CNN or LSTM models on benchmark turbofan datasets.

The hybrid prognostic framework proposed by Y. Li *et al*. (2023) for estimating the remaining useful life of turbofan engines has a strong point in integrating physics-based features and neural networks. However, it does not solve the problem of real-time accurate fault identification. The diagnostic strategy applied here minimises classification errors under unstable and incomplete input data. W. Tang *et al*. (2020) developed a PHM approach for marine hybrid energy systems focused on battery lifetime prediction and optimisation of diesel-electric components. The lack of diagnostic error analysis and quantitative validation limits the applicability of the method for evaluating classification accuracy. In contrast, the present study implements a formalised approach to configuration selection based on metric comparison under various operating scenarios. Thus, the presented results confirm the effectiveness of the developed configuration for diagnosing SPP failures under uncertainty, data limitations, and variable loads. The comparison with current research emphasises the competitiveness of the approach not only in terms of accuracy but also in its versatility for engineering diagnostics of complex technical systems.

Conclusions

The conducted study quantitatively evaluated the diagnostic accuracy of three methodological configurations for SPP fault detection: the basic CBR model, the hybrid CBR combined with probabilistic analysis (BNs and MMs), and the integrated configuration including simulation modelling of cascading failures. The results demonstrate that the CBR approach ensures stable and interpretable diagnostics under nominal operating conditions, with average accuracy around 82-85%. Its key advantage is simplicity and low computational demand, while the main limitation is sensitivity to data incompleteness and noise. The probabilistic configuration (CBR + BNs/MMs) improves fault sensitivity and diagnostic stability, increasing accuracy to 88-90%. It effectively handles uncertainty and gradual degradation but requires accurate prior probabilities and has moderate computational complexity. The fully integrated configuration (CBR + BNs/MMs + Simulation) achieved the highest performance – up to 93-94% accuracy and a minimum diagnostic error of 6%. It demonstrated robustness to noise, incomplete data, and load fluctuations, providing the most balanced and reliable results across all operational scenarios.

Analysis of confusion matrices revealed that the main diagnostic errors in isolated models are false negatives (missed pre-failure states), while integrated methods significantly reduce this error type – by approximately 35% compared to CBR alone. Visualisation of classification errors confirmed that the integration of probabilistic and simulation components enhances sensitivity without increasing false alarms, ensuring consistent performance in complex operating conditions. The optimal weighting of diagnostic components was determined as $\alpha = 0.6$ (CBR), $\beta = 0.2$ (probabilistic models), and $\gamma = 0.2$ (simulation modelling), which minimises total error and ensures model adaptability under variable operational environments.

In practical terms, simplified CBR-based diagnostics are sufficient for stable, nominal operation, whereas under high load and information scarcity, the integrated multimethod configuration is preferable. Future research should focus on expanding the adaptive calibration of model weights using real-time learning mechanisms and integrating digital twin technologies for continuous system monitoring and predictive maintenance in marine and industrial applications.

Acknowledgements

None.

Funding

The study was not funded.

Conflict of Interest

None.

References

[1] Aburakhia, S., Myers, R., & Shami, A. (2022). A hybrid method for condition monitoring and fault diagnosis of rolling bearings with low system delay. *IEEE Transactions on Instrumentation and Measurement*, 71, article number 3519913. doi: 10.1109/TIM.2022.3198477.

[2] Brito, L.C., Susto, G.A., Brito, J.N., & Duarte, M.A.V. (2021). An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery. *ArXiv*. doi: 10.48550/arXiv.2102.11848.

[3] Cui, Y., Sun, Y., Cui, J., Zhao, F., & Yuan, M. (2025). Digital twin for marine diesel engine: Enhancing predictive maintenance and operational efficiency. In *CSAA/IET international conference on aircraft utility systems (AUS 2024)*. Xi'an: CSAA/IET.

[4] Daya, A.A., & Lazakis, I. (2023). Component criticality analysis for improved ship machinery reliability. *Machines*, 11, article number 737. doi: 10.3390/machines11070737.

[5] Fera, F., & Spandonidis, C. (2024). A fault diagnosis approach utilizing artificial intelligence for maritime power systems within an integrated digital twin framework. *Applied Sciences*, 14, article number 8107. doi: 10.3390/app14188107.

[6] Hasan, A., Asfihani, T., Osen, O., & Bye, R.T. (2024). Leveraging digital twins for fault diagnosis in autonomous ships. *Ocean Engineering*, 292, article number 116546. doi: 10.1016/j.oceaneng.2023.116546.

[7] Li, Y., Chen, Y., Hu, Z., & Zhang, H. (2023). Remaining useful life prediction of aero-engine enabled by fusing knowledge and deep learning models. *Reliability Engineering & System Safety*, 229, article number 108869. doi: 10.1016/j.ress.2022.108869.

[8] Lu, C., Tao, L., Ma, J., Cheng, Y., & Ding, Y. (2024). *Fault diagnosis and prognostics based on cognitive computing and geometric space transformation*. New York: Springer. doi: 10.1007/978-981-99-8917-1.

[9] Moon, H., Choi, J., & Cha, S. (2021). A multi-state Markov model to infer the latent deterioration process from the maintenance effect on reliability engineering of ships. *ArXiv*. doi: 10.48550/arXiv.2111.14368.

[10] Morato, P.G., Papakonstantinou, K.G., Andriotis, C.P., Nielsen, J.S., & Rigo, P. (2020). Optimal inspection and maintenance planning for deteriorating structural components through dynamic Bayesian networks and Markov decision processes. *Structural Safety*, 94, article number 102140. doi: 10.1016/j.strusafe.2021.102140.

[11] Nguyen, B.L.H., Vu, T.V., Nguyen, T.-T., Panwar, M., & Hovsapian, R. (2023). Spatial-temporal recurrent graph neural networks for fault diagnostics in power distribution systems. *IEEE Access*, 11, 46039-46050. doi: 10.1109/ACCESS.2023.3273292.

[12] Orhan, M., & Celik, M. (2024). A literature review and future research agenda on fault detection and diagnosis studies in marine machinery systems. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, 238(1), 3-21. doi: 10.1177/14750902221149291.

[13] Quality-One International. (n.d.). *Fault Tree Analysis (FTA)*. Retrieved from https://quality-one.com/fta.

[14] Panagiotopoulou, V., Petriconi, E., Giglio, M., & Sbarufatti, C. (2025). Deep learning-based identification of shaft imbalance faults in rotating machinery using the NARX model. *Journal of Vibration Engineering & Technologies,* 13, article number 261. doi: 10.1007/s42417-025-01823-8.

[15] Rigas, S., Tzouveli, P., & Kollias, S. (2024). An end-to-end deep learning framework for fault detection in marine machinery. *Sensors*, 24(16), article number 5310. doi: 10.3390/s24165310.

[16] Tang, W., Roman, D., Dickie, R., Robu, V., & Flynn, D. (2020). Prognostics and health management for the optimization of marine hybrid energy systems. *Energies*, 13(18), article number 4676. doi: 10.3390/en13184676.

[17] Ture, B.A., Akbulut, A.H., Zaim, A.H., & Çatal, C. (2023). Stacking-based ensemble learning for remaining useful life estimation. *Soft Computing,* 28, 1337-1349. doi: 10.1007/s00500-023-08322-6.

[18] Vychuzhanin, V., & Vychuzhanin, A. (2025). *Intelligent diagnostics of ship power plants: Integration of case-based reasoning, probabilistic models, and ChatGPT. A universal approach to fault diagnosis and prognostics in complex technical systems*. Lviv-Torun: Liha-Pres. doi: 10.36059/978-966-397-516-0.

[19] Vychuzhanin, V., Rudnichenko, N., Boyko, V., Shibaeva, N., & Konovalov, S. (2016). Devising a method for the estimation and prediction of technical condition of ship complex systems. *Eastern-European Journal of Enterprise Technologies*, 84(6/9), 4-11. doi: 10.15587/1729-4061.2016.85605.

[20] Xia, T., Han, D., Jiang, Y., Shao, Y., Wang, D., Pan, E., & Xi, L. (2024). Remaining useful life estimation based on selective ensemble of deep neural networks with diversity. *Advanced Engineering Informatics*, 62, article number 102608. doi: 10.1016/j.aei.2024.102608.

[21] Zhang, Y., Zhu, W., Jin, C., & Liang, Z. (2022). Hybrid modeling and simulation for shipboard power system considering high-power pulse loads integration. *Journal of Marine Science and Engineering*, 10, article number 1507. doi: 10.3390/jmse10101507.

Трисценарний аналіз точності діагностики несправностей у складних технічних системах

Володимир Вичужанін

Доктор технічних наук, професор
Національний університет «Одеська політехніка»
65044, просп. Шевченка, 1, м. Одеса, Україна
https://orcid.org/0000-0002-6302-1832

Олексій Вичужанін

Доктор філософії, асистент
Національний університет «Одеська політехніка»»
65044, просп. Шевченка, 1, м. Одеса, Україна
https://orcid.org/0000-0001-8779-2503

Анотація. Метою цього дослідження було проведення трисценарного порівняльного аналізу точності інтелектуальної діагностики несправностей у складних технічних системах на прикладі суднових енергетичних установок (СЕУ). Дослідження було спрямоване на визначення конфігурації діагностичних методів, яка забезпечує найвищу точність і робастність за різних експлуатаційних умов. Розглянуто три методологічні конфігурації: базову модель на основі Case-Based Reasoning (CBR); CBR, доповнену ймовірнісним аналізом із використанням баєсівських мереж і марковських ланцюгів; а також інтегровану модель, що поєднує CBR, ймовірнісні методи та імітаційне моделювання каскадних відмов. Експерименти проведено для трьох типових сценаріїв експлуатації – номінального режиму, режиму високого навантаження та режиму з обмеженими діагностичними даними, що відображають реальні умови морської експлуатації. Для оцінювання ефективності застосовано стандартні метрики класифікації: Accuracy, Recall і F1-score. Результати показали, що базова конфігурація CBR забезпечує середню точність 82–85 % за номінальних умов, проте суттєво втрачає ефективність за неповних даних. Інтеграція з ймовірнісними моделями підвищує стабільність метрик, збільшуючи точність до 88–90 %. За оптимального розподілу вагових коефіцієнтів (CBR – αd = 0,6, ймовірнісні моделі – βd = 0,2, імітаційне моделювання – γd = 0,2) досягається мінімальна діагностична похибка – 6 %, а загальна точність перевищує 93 %, навіть за наявності шумів і пропусків даних. Аналіз матриць плутанини та візуалізацій похибок показав, що інтегровані конфігурації зменшують кількість помилок другого роду приблизно на 35 % порівняно з ізольованими підходами. Тривимірні графіки залежності точності від вагових коефіцієнтів підтверджують стійкий максимум у зоні збалансованих параметрів і підкреслюють значущість імітаційного компонента за складних експлуатаційних умов. Отримані результати дали змогу сформулювати практичні рекомендації щодо вибору діагностичних конфігурацій: CBR + баєсівські мережі доцільно застосовувати за стабільних режимів роботи, а повну інтеграцію всіх компонентів – за умов перевантаження та дефіциту інформації. Запропонована методологія може бути адаптована до інших інтелектуальних діагностичних систем, що працюють за умов невизначеності, змінного навантаження та неповних даних, зокрема у кіберфізичних та промислових системах. Представлений підхід є універсальним і масштабованим рішенням для прикладних задач діагностики, які потребують високої точності, адаптивності та стійкості

Ключові слова: Case-Based Reasoning; баєсівські мережі; марковські ланцюги; імітаційне моделювання; діагностичні метрики; невизначеність сенсорів, адаптивна підтримка прийняття рішень

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 23.07.2025 Revised: 28.10.2025 Accepted: 23.12.2025

UDC 621.396.97 DOI: 10.31649/vitce/3.2025.41

Intelligent frequency management in FANET: Fuzzy logic/routing and adaptive frequency hopping

Roman Zaivyi*

Postgraduate Student
Lviv Polytechnic National University
79000, 12 Stepan Bandera Str., Lviv, Ukraine
https://orcid.org/0009-0003-4096-4111

Volodymyr Pavlysh

PhD in Technical Sciences, Professor
Lviv Polytechnic National University
79000, 12 Stepan Bandera Str., Lviv, Ukraine
https://orcid.org/0009-0004-3996-5923

Abstract. The study aimed to experimentally evaluate the effectiveness of intelligent frequency management in swarm networks of unmanned aerial vehicles (UAVs) using fuzzy logic and adaptive frequency hopping. The object of analysis was three frequency control methods: fixed frequency, classical frequency hopping, and the proposed adaptive method, which combines fuzzy logic decision making with context-dependent routing. The research was conducted in a MATLAB R2024a and Python 3.12 simulation environment on a model of five UAVs moving within an area of 1000×1000 m, incorporating changes in topology, signal level, signal-to-noise ratio, and energy characteristics of the nodes. The results demonstrated that the developed adaptive method provides the highest communication efficiency among the approaches studied. The packet delivery rate remained at 0.93-0.95 even in the presence of narrowband interference, which is 25-30% higher than the basic methods. The average end-to-end transmission delay decreased to 43 ms compared to 61 ms in the classic frequency hopping scheme and 78 ms in fixed mode. Power consumption decreased by 12-19%, and the average switching frequency was halved (≈ 2 times/s compared to 4.2 times/s in the classic mode), which indicates the optimisation of the controller's operation. Statistical analysis confirmed the significant impact of the method type on all key communication performance indicators ($p < 0.05$), which confirms the reliability of the results obtained and the reproducibility of the system in a series of simulation experiments. The proposed approach provides autonomous optimisation of data transmission routes and maintenance of a stable communication channel even in dynamic environments, which creates prospects for the development of a new generation of intelligent UAV networks focused on real-time monitoring, reconnaissance and coordination tasks. The research results can be used by developers of unmanned systems, communications engineers and network technology specialists to create more interference-resistant, energy-efficient and self-learning communication systems

Keywords: unmanned aerial vehicles; swarm networks; radio module; energy efficiency; packet delivery ratio; node mobility

Introduction

With the development of swarm unmanned aerial vehicle (UAV) technologies, the issue of secure communication in dynamically changing topologies, limited energy resources, and radio interference has become one of the central topics in modern telecommunications research. Flying Ad Hoc Networks (FANETs) are considered a key infrastructure for collective drone interaction in monitoring, reconnaissance, search and rescue missions, and military operations. In contrast to ground-based mobile ad hoc networks (MANET), swarm networks are characterised by increased node mobility and the complexity of maintaining reliable communication channels. In such conditions, classic fixed frequency access schemes become ineffective, and traditional routing methods are unable to respond in a timely

Suggested Citation:
Zaivyi, R., & Pavlysh, V. (2025). Intelligent frequency management in FANET: Fuzzy logic/routing and adaptive frequency hopping. *Information Technologies and Computer Engineering*, 22(3), 41-53. doi: 10.31649/vitce/3.2025.41

*Corresponding author

manner to changes in air conditions and node energy levels. Therefore, the concepts of cognitive radio, fuzzy logic, and adaptive frequency hopping, which provide intelligent real-time radio resource management, are becoming increasingly relevant.

S. Semendiai (2023) noted that the use of cognitive radio in conditions of active use of electronic warfare means can dynamically adapt transmission parameters and avoid noisy frequencies. This approach significantly increases the noise immunity of the system and the efficiency of spectrum use. An analysis of the FANET architecture conducted by S. Valuyskyi & O. Ponomarenko (2025) showed that the main challenges remain network connectivity support, load balancing, and the lack of coordination between channel control and routing levels. The study emphasised that traditional protocols borrowed from MANET do not incorporate the high dynamics of UAV movement and the variability of air parameters, which causes route breaks and packet loss. In this context, fuzzy logic is considered an effective decision-making tool in conditions of uncertainty regarding channel characteristics and the energy state of nodes. A similar idea was developed by R. Bieliakov & O. Fesenko (2023), proposing a model of intelligent resource management in MANET, combining the analysis of signal quality parameters, load and energy constraints. Their concept proves the promise of integrating fuzzy algorithms and adaptive resource management into unmanned communication systems, which creates the basis for further optimisation of the FANET architecture.

A significant contribution to the development of intelligent routing methods was made by A. Rahmani *et al.* (2022), proposing an improved Optimised Link State Routing Plus protocol, implemented based on fuzzy logic inference. The system evaluates channel quality parameters, node residual energy, and neighbourhood density, forming a flexible weighting function for selecting the optimal repeater. This approach significantly reduced the number of retransmissions and increased the average packet delivery rate. Continuing their research, M. Prakash *et al.* (2024) proposed a reinforcement learning routing scheme, where the agent gradually learns to optimise path selection, covering the history of topology changes, noise level, and residual node energy. Such a hybrid system combines the intellectual capabilities of probabilistic models with the adaptability of fuzzy logic. Another direction is demonstrated by M. Hosseinzadeh *et al.* (2024), developing an energy-oriented routing algorithm with local data filtering. A distinctive feature of the algorithm is the preliminary assessment of "reception quality" based on the combined characteristics of the Signal-to-Noise Ratio (SNR) and load, which reduces the number of unnecessary retransmissions and thus extends the autonomous operation time of the nodes. A similar approach was developed by S. Khan *et al.* (2022), proposing the Ant-HocNet protocol, based on optimised fuzzy logic for UAV swarm networks in FANET. Their model combines a bio-inspired ant routing algorithm with an adaptive fuzzy controller that dynamically evaluates the energy balance of nodes, the degree of connectivity, and the intensity of interference. This hybrid approach has improved packet delivery rates and reduced transmission delays compared to traditional Optimised Link State Routing and Ad hoc On-Demand Distance Vector methods, while increasing route stability in a dynamic swarm environment. In the context of generalising architectural and protocol solutions for FANET, the conclusions of T.K. Bhatia *et al.* (2024) are substantial, emphasising the critical role of frequency adaptation, predictive routing, and intelligent spectrum management mechanisms in ensuring the stability of swarm networks under conditions of high mobility and radio interference. The study emphasised that it is hybrid approaches combining bio-inspired algorithms, fuzzy logic and frequency dynamics that demonstrate the highest efficiency in real-time scenarios.

Scientists emphasised dynamic frequency hopping methods. Thus, C. Atheeq *et al.* (2024) developed a chaotically controlled frequency hopping mechanism that uses random hopping maps with mathematically guaranteed unpredictability, thanks to which the system demonstrates high resistance to directional jamming and frequency-selective interference. This approach is particularly effective in electronic warfare scenarios. At the same time, J. Alotaibi (2025) noted that predictive routing with a built-in fuzzy frequency controller provides higher throughput and lower power consumption compared to classical Frequency Hopping Spread Spectrum (FHSS) methods. Thus, the combination of adaptive frequency hopping, fuzzy logic, and route optimisation forms the basis for the creation of intelligent frequency control systems in UAV swarm networks.

A generalised analysis of the literature showed that integrating fuzzy logic into dynamic frequency selection mechanisms can simultaneously address channel quality, node energy status, and routing requirements. However, there is still a lack of comprehensive models capable of ensuring the coordinated operation of these components in real time. In this context, the study aimed to experimentally verify the effectiveness of an intelligent frequency control system in UAV swarm networks, based on the principles of fuzzy logic and adaptive frequency hopping. The main objectives of the study were to develop a simulation model of FANET, create a fuzzy frequency control controller, conduct a comparative analysis of the effectiveness of three frequency control modes, and perform a further statistical verification of the results obtained.

Materials and Methods

The material basis of the study was the MATLAB R2024a and Python 3.12 software packages used to model the operation of FANET. The mathematical model of the network was constructed as a discrete-event system focused on reproducing changes in node positions, topology dynamics, communication channel parameters, and the influence of external interference. The simulation environment contained five UAVs operating in a square area of 1000×1000 m. The model simulated generalised tactical-class quadcopters

with FANET-typical manoeuvrability and speed characteristics, which made it possible to reproduce the topology dynamics characteristic of swarm networks without reference to a specific platform. The movement of the drones was implemented using the Random Waypoint model at a speed of 0-10 m/s, which created the topology variability characteristic of FANET. The flight was modelled at a fixed altitude of 100 m above ground level, which corresponds to the typical conditions of use of tactical quadcopters in open space. Isotropic antennas with the same radiation pattern were used for radio communication, which ensured uniform coverage in the horizontal plane and eliminated the influence of directivity on channel quality.

The radio modules operated in the 2.4 GHz band with the ability to switch between three discrete channels. The transmitter power was 20 dBm, and the receiver sensitivity was -90 dBm with a minimum SNR ⩾ 6 dB. The network functioned as a multi-hop data transmission system: each UAV could act as an intermediate repeater for other nodes, ensuring stable communication under conditions of constantly changing topology. The Ad hoc On-Demand Distance Vector protocol was used for routing, which forms routes as needed and dynamically updates them when the topology changes, which is a typical and effective approach for mobile ad hoc networks with frequent connection breaks, in particular, FANET. The model also considered the spatial and topological characteristics of the connection, determined by the relative positions of the UAVs and the dynamics of their movement. For each moment in time, inter-node distances, the degree of network graph connectivity, the presence or absence of alternative multi-hop routes, and changes in topological components (cluster formation and disintegration) were calculated. These parameters were used to assess routing stability and influenced the decision of the fuzzy controller to switch frequencies, since the nature of the topology determines the level of interference, the availability of neighbouring nodes, and the probability of packet loss.

The channel model was based on the free-space attenuation law with the addition of additive white Gaussian noise. To reproduce the conditions of radio-electronic interference on one of the channels, narrowband interference equivalent to the source signal at approximately 100 m was introduced. The source of radio interference was modelled as a stationary transmitter with a radiation power of 15 dBm, a bandwidth of 200 kHz and a fixed location within the test area, which created stable spectral pressure in a given frequency range without changing the spatial position. This was used to evaluate the impact of localised narrowband interference on FANET communication parameters in different frequency control modes.

Data was transmitted in the form of 512-byte UDP packets at a rate of 10 packets per second. Each simulation session lasted 600 seconds, and each scenario was repeated 30 times to average the results. The network architecture included a coordinator drone responsible for synchronising frequency hopping between all FANET nodes, which ensured consistent channel control.

The methodological part of the study included the development of an intelligent frequency control controller based on fuzzy logic by E.H. Mamdani & S. Assilian (1975), a series of simulations in different frequency control modes, and statistical processing of the results. At the first methodological stage, a fuzzy controller was created, the input variables of which were Received Signal Strength Indicator (RSSI), SNR, residual node energy level (E) and load (Q), which characterised the intensity of route traffic. Sigmoid functions were used for RSSI and SNR, and triangular functions were used for E and Q for phasing. The controller's knowledge base consisted of 18 production rules of the "if – then" type. The output variable was the assessment of the need for frequency switching; defasification was performed using the average maximum method. To prevent excessive frequency fluctuations, time hysteresis was applied: the minimum interval between two switches was 0.5 s.

The second stage involved modelling three frequency control modes: fixed frequency operation, FHSS, and adaptive frequency hopping with fuzzy logic. All nodes operated in synchronous frequency mode under the control of a coordinator. In each scenario, the same motion trajectories, initial conditions, and random number generators were reproduced, ensuring the reproducibility of the experiments. The starting positions of the UAVs were determined randomly with a uniform distribution within a 1,000×1,000 m area, after which the drones moved according to the Random Waypoint model with an initial pause of 1 s and a minimum stay at the selected point of 2 s before changing direction. A separate component of the study was a comparative analysis of the effectiveness of three frequency control approaches: fixed frequency, FHSS, and Adaptive FH, aimed to determine their differences in terms of connection stability, transmission delay, energy consumption, and switching frequency.

At the final stage, statistical data processing was performed in Python 3.12. The normality of distributions was checked using the Shapiro-Wilk test, and the homogeneity of variances was checked using the Levene test. To assess inter-mode differences, a two-factor ANOVA ("method type × distance between nodes") was used. In cases of violation of homogeneity of variances, Welch's criterion with Bonferroni correction was used. The significance level was $p < 0.05$. The effect size was determined using Cohen's d or Glass' Δ, depending on the nature of the variances. Additionally, a normalised multi-criteria analysis was performed to summarise the performance of the methods according to four key indicators: Packet Delivery Ratio (PDR), Delay, Energy Consumption, and Switching Rate. To ensure the comparability of heterogeneous values, min-max normalisation was applied, where each indicator was converted to a dimensionless scale [0;1] by dividing it by its maximum (for PDR) or minimum (for Delay, Energy and Switching Rate) value within the analysed methods. This formulated an integral representation of efficiency and constructed multi-criteria diagrams, which were used for comparative interpretation of the results.

Results

Parameters and structure of the FANET model

Analysis of the constructed model showed that the efficiency of the swarm network is largely determined by the frequency control architecture. The simulation confirmed that the use of a coordinator scheme with local decision-making modules provides an optimal combination of node autonomy and synchronisation of their actions during frequency hopping. The structure of the model shown in Figure 1 demonstrates that each FANET node functions as an autonomous element with its own radio frequency assessment module, while the coordinator is responsible for distributing global commands for channel switching. Simulation results have shown that this organisation minimises coordination delays and maintains routing stability even in the presence of narrowband interference in part of the spectrum. Compared to uncoordinated models, synchronous channel hopping reduced the probability of desynchronisation and packet loss, which was particularly evident in scenarios with sharp changes in SNR.

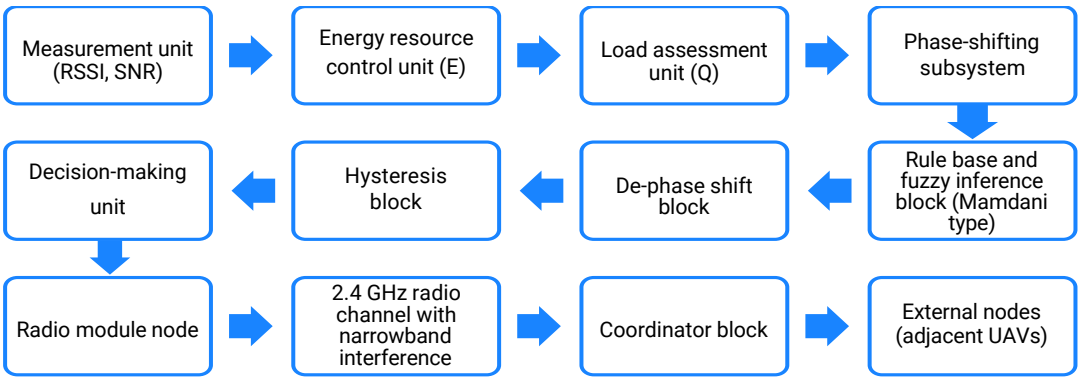


Figure 1. Structural diagram of the frequency control system in the FANET network node
Source: compiled by the authors based on the parameters of the simulation environment MATLAB R2024a and Python 3.12

As shown in Figure 1, the information flow begins with the RSSI and SNR, E, and Q estimation blocks. The obtained parameters are sent to the phasing subsystem, where they are converted into linguistic variables for further processing in the E-type fuzzy inference block. E.H. Mamdani & S. Assilian (1975). The controller's rule base contains 18 heuristic "if – then" constructions that describe the relationships between channel quality, energy level, and node load. The result of fuzzy inference is sent to the defuzzification block, which forms a numerical value η – degree of need for frequency switching. To avoid excessive fluctuations, a hysteresis block with a time threshold of at least 0.5 s is used, after which the signal is sent to the decision-making block. If $\eta > 0.5$, the radio module executes the command to change the operating channel. Switching is conducted within the 2.4 GHz range and is coordinated with the coordinator block, which is responsible for synchronising all UAVs during transitions between frequencies. The communication channel is modelled based on additive white Gaussian noise and possible narrowband interference, which can be used to assess the system's resistance to external interference. As a result, the proposed structure implements a complete closed-loop frequency control circuit from data collection to adaptive hopping between channels.

The key numerical parameters used in the simulations are shown in Table 1. They determine the spatial and temporal characteristics of the swarm network, the technical parameters of the radio modules, the signal update frequency, and the total duration of the experiment. The selected values are consistent with typical conditions for testing communication systems based on ad hoc communication protocols, which ensures the reproducibility of results.

Table 1. Basic parameters for modelling the FANET network

Parameter	Value
Number of nodes	5
Region size	1,000×1,000 m
Node movement model	Random Waypoint (0-10 m/s)
Transmitter power	20 dBm
Receiver sensitivity	-90 dBm (at SNR ≥ 6 dB)
Number of channels	3 (2.4 GHz range)
Channel assessment interval	0.5 s
Maximum jump frequency	2 times/s (adaptive method)
FHSS rate (standard)	10 times/s
Simulation duration	600 s (10 min)

Source: compiled by the authors based on the parameters of the simulation environment MATLAB R2024a and Python 3.12

The selected parameters provide an optimal balance between realism and controllability of the simulation. The speed range of 0-10 m/s can reproduce the behaviour of UAVs in real conditions, and the channel evaluation period of 0.5 s ensures a sufficient frequency of updating information about the quality of communication. The maximum hopping frequency limit (2 times/s) prevents excessive load on the network stack and supports the use of the Adaptive FH algorithm to maintain stable operation in dynamic topologies. For comparison, the basic Fixed and FHSS methods are used as reference scenarios to demonstrate the difference in adaptation speed and resistance to external interference. Thus, the selected simulation configuration reproduces typical conditions of swarm interaction between UAVs and quantifies the effectiveness of the proposed intelligent adaptive frequency hopping mechanism.

Implementation and algorithmic principles of a fuzzy frequency control controller in FANET

Within the scope of the study, an intelligent frequency control controller was implemented, which operates on the principle of fuzzy logic inference of the type described by E.H Mamdani & S. Assilian (1975). Such a controller can autonomously decide on the need to change the working communication channel depending on the state of the airwaves, energy resources, and node load. The main input parameters of the system are RSSI, SNR, E, and Q (Table 2). These parameters reflect the actual state of communication and can be used for the assessment of the suitability of the selected channel for further operation. Sigmoidal membership functions are used for RSSI and SNR, as these indicators change smoothly under the influence of noise. For energy and load characteristics (E and Q), triangular functions are used, which ensure a smooth transition between states.

Table 2. Characteristics of input variables of a fuzzy controller

Parameter	Type of membership function	Linguistic levels	Value range
RSSI	Sigmoid	Low (-90...-75 dBm), medium (-75...-55 dBm), high (-55...-40 dBm)	From -90 to -40 dBm
SNR	Sigmoid	Low (0...10 dBm), high (10...30 dBm)	From 0 to 30 dB
E	Triangular	Low (0.0...0.3), medium (0.3...0.7), high (0.7...1.0)	From 0 to 1
Q	Triangular	Weak (0.0...0.5), substantial (0.5...1.0)	From 0 to 1

Source: compiled by the authors based on MATLAB R2024a and Python 3.12 modelling

Following Table 2, the fuzzy controller is based on four input parameters (RSSI, SNR, E and Q) and one output variable η. Three linguistic states are used for RSSI ("low", "medium", "high"), two for SNR ("low", "high"), and two for energy and load ("low"/"high" and "weak"/"substantial", respectively). This configuration provides sufficient sensitivity to changes in channel quality and network conditions without overly complicating the rule base. Sigmoidal membership functions for RSSI and SNR can be used for smooth tracking of signal degradation in noisy conditions, while triangular functions for E and Q separated "energy saving" and "high load" modes. The output variable η is normalised in the range [0;1] and interpreted as the degree of need for frequency switching, which simplifies threshold decision-making. Together, this indicates that the controller is focused not only on maintaining channel quality but also on conserving energy resources and ensuring stable node operation during intensive routing.

The fuzzy controller knowledge base consists of 18 production rules in the "if – then" format, which determine the system's response to various combinations of input parameters and control the frequency selection. They cover typical situations that arise during the operation of a mesh network, for example, when the signal level is low, the signal-to-noise ratio decreases, but the node's energy resources are still sufficient, and the load is significant, the controller decides to switch to another channel to avoid network overload. On the other hand, in cases of a stable signal, high SNR and reduced energy level, it is advisable to remain at the current frequency to reduce energy consumption. Each of the rules forms a partial recommendation for action, after which the system calculates an integral indicator of the degree of need for a frequency change. If this value exceeds the threshold level of 0.5, the controller automatically initiates a command to change the communication channel (Mamdani & Assilian, 1975). To avoid excessive switching, a hysteresis mechanism is implemented, with a time delay of at least 0.5 seconds between successive transitions, which prevents the effect of unstable oscillations ("flip-flop"). As a result, the controller operates cyclically: first, data is collected, then phasing, fuzzy inference, defasing, hysteresis checking and, if necessary, frequency switching. This sequence ensures the adaptability and autonomy of each FANET network node, maintaining communication stability and high data transmission efficiency in changing environmental conditions.

Package delivery dynamics and communication stability with adaptive frequency control in FANET

A comparative analysis of the effectiveness of three approaches – fixed frequency, FHSS, and Adaptive FH – showed significant differences in connection stability and the number of lost packets over time (Fig. 2). At the beginning of the experiment (100 s), all methods showed similar results, but as the simulation duration increased, the difference became more pronounced. In the classic FHSS mode, PDR remained stable up to 300 s, but after the appearance of narrowband interference, the indicator gradually decreased. For a fixed frequency (Fixed), the degradation of communication occurred faster because the system

did not have mechanisms to avoid interference. In contrast, the adaptive method with a fuzzy controller maintained high PDR values even in difficult air conditions thanks to timely switching to clean channels.

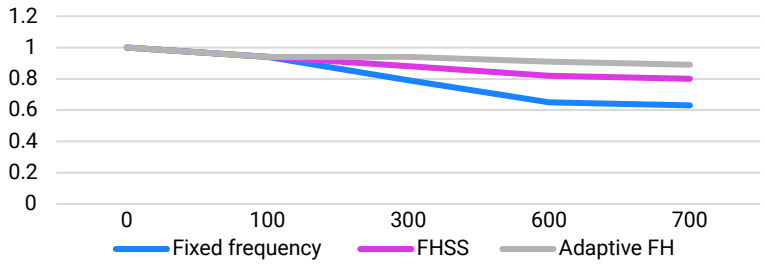


Figure 2. PDR dynamics over time
Source: compiled by the authors based on MATLAB R2024a and Python 3.12 simulation

The dynamics of the packet delivery ratio demonstrate a steady increase in the efficiency of the proposed adaptive method compared to the basic fixed frequency and FHSS schemes. In the first 100 seconds, all methods have similar values (0.94-0.97), but over time, the difference increases significantly. At the 300-second mark, the adaptive scheme maintains PDR ≈ 0.94, while FHSS drops to 0.88 and Fixed to 0.79. By the end of the simulation (600 seconds), the difference is most pronounced: Adaptive FH maintains a delivery rate of over 0.9, while fixed frequency drops to 0.65. This behaviour confirms that the proposed method can compensate for the influence of narrowband interference and maintaining a stable communication channel in a dynamic FANET environment. The implementation of an intelligent adaptive frequency hopping controller significantly improves the reliability of data transmission in a FANET swarm network. Contrary to traditional Fixed Frequency and FHSS approaches, the proposed system provides a dynamic response to changes in the air condition, minimising packet loss and preventing channel overload. An adaptive mechanism based on fuzzy logic can be used in each node to switch to less noisy frequencies in a timely manner, maintaining high network throughput and communication stability throughout the simulation. The results confirm the effectiveness of integrating fuzzy control into FANET frequency algorithms and lay the foundation for further system optimisation, considering energy and topological constraints.

Comparison of Fixed, FHSS and Adaptive FH in FANET: Spatial and energy aspects

A study of the effect of distance between nodes in a swarm network showed that as the distance increases, PDR gradually decreases for all methods considered, but the rate of decline varies significantly (Fig. 3). At short distances (up to 200 m), all three schemes (fixed frequency, FHSS, and Adaptive FH) demonstrated almost identical efficiency with PDR of about 0.97-0.99. However, with an increase in the spatial gap to 400-500 m, the stability of the connection was significantly higher for the adaptive method.

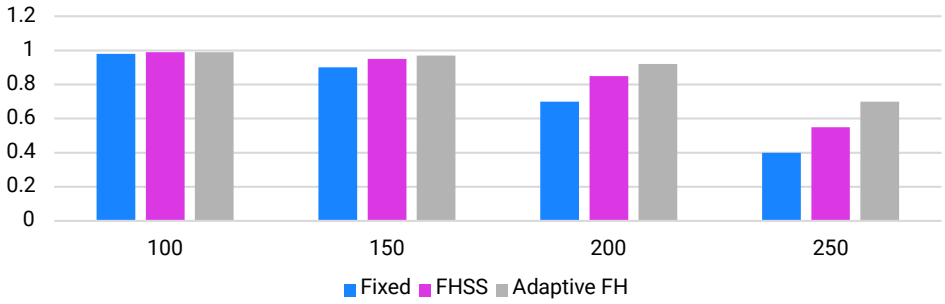


Figure 3. PDR dependence on the distance between nodes for Fixed, FHSS, and Adaptive FH
Source: compiled by the authors based on MATLAB R2024a and Python 3.12 simulation

As shown in the diagram, while Fixed Frequency demonstrates packet loss of up to 60% over long distances, Adaptive FH provides stable transmission quality, maintaining a PDR value of no less than 0.70 even when the inter-node distance increases. This means a 25-30% expansion of the communication coverage area, thanks to dynamic frequency hopping and the use of a fuzzy controller that selects the least noisy channels. In addition to increased resistance to interference, the proposed method demonstrated better energy efficiency. Energy consumption analysis showed that Adaptive FH reduces the total energy consumption of nodes by approximately 12% compared to FHSS and by almost 20% compared to fixed frequency. This is due to the optimisation of the hopping frequency: the controller initiates a channel change only when there is a significant deterioration in signal quality,

whereas classic FHSS performs hopping periodically, regardless of the channel status. As a result, the average switching frequency for Adaptive FH was about 2 times/s, which is half that of FHSS and did not affect the stability of data transmission. In addition, the average end-to-end delay was evaluated. A normalised comparison of four key indicators – PDR, end-to-end delay, power consumption, and switching frequency – confirmed the clear advantage of adaptive frequency control over other methods. After converting the values to a dimensionless scale [0;1], the study determined that Adaptive FH demonstrates the highest normalised packet delivery rate and the lowest delay and power consumption values, while maintaining a moderate switching frequency. Compared to FHSS, the adaptive approach provided a significantly better balance between channel stability and energy efficiency, while fixed frequency, despite the minimum number of switches, was significantly inferior in terms of connection stability. Table 3 shows the comparative results of the effectiveness of the three frequency control methods, fixed frequency, FHSS and Adaptive FH, based on three key indicators.

Table 3. Integral performance indicators for Fixed, FHSS and Adaptive FH methods

Method	Energy consumption (mJ/cycle)	Relative energy consumption, % (from Fixed = 100%)	Average switching frequency (times/second)	Relative switching frequency, % (from Fixed = 100%)	Average transmission delay (ms)	Relative delay, % (from Fixed = 100%)
Fixed frequency	8.4	100	0.3	100	78	100
FHSS	7.7	91.7	4.2	1,400	61	78.2
Adaptive FH	6.8	81	2	666.7	43	55.1

Source: compiled by the authors based on MATLAB R2024a and Python 3.12 simulation

The data obtained shows that the adaptive method not only consumes less energy but also reduces the switching frequency by more than half compared to classic FHSS, which proves the effective use of the spectrum. The average end-to-end transmission delay is also the lowest for Adaptive FH (≈ 43 ms), while FHSS reaches 61 ms and fixed frequency reaches 78 ms, which is due to fewer retransmissions and no unnecessary switching. This proves that intelligent frequency shifting not only maintains high communication quality but also minimises delay without reducing PDR. The adaptive mechanism, based on a fuzzy logic controller, provides a balanced relationship between stability, energy efficiency and speed. It ensures an adaptive response of the FANET network to channel state changes, extends the spatial coverage, reduces the load on nodes, and improves the overall efficiency of the data transmission system. The results, visualised in the form of a multi-criteria diagram, show that Adaptive FH occupies the largest area of the normalised profile, which effectively indicates its integral advantage over other methods in FANET with variable topology. Figure 4 shows a generalised comparative diagram of the efficiency of the three approaches: fixed frequency, FHSS and Adaptive FH. For ease of interpretation, the indicators that characterise performance positively (PDR) are normalised according to the "more is better" principle, and those that reflect losses or delays (Delay, Energy, Switching Rate) are normalised according to the "less is better" principle.

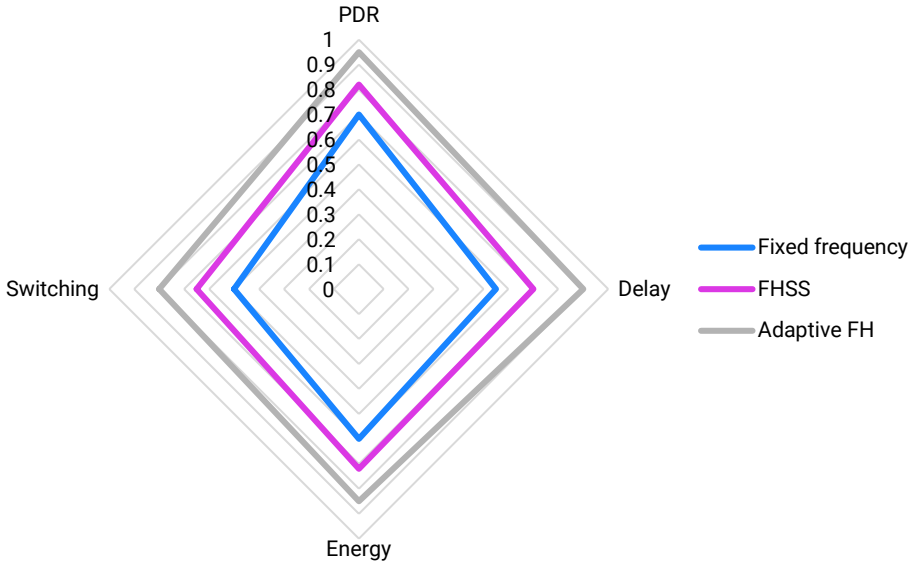


Figure 4. Generalised comparative diagram of method effectiveness
Source: compiled by the authors based on normalised analysis in MATLAB R2024a and Python 3.12

The proposed adaptive method based on fuzzy control demonstrates the most balanced profile among all the analysed schemes. Its normalised PDR and Delay indicators are close to the maximum (0.93-0.95), which indicates stable communication and minimal time delays. At the same time, power consumption and switching frequency remain the lowest, at 0.80 and 0.78, respectively, relative to the baseline methods, indicating high energy efficiency and optimal use of frequency resources. A comparative assessment shows that traditional fixed frequency has a narrow profile due to high energy consumption and frequent packet loss as the distance between nodes increases. Classic FHSS demonstrates some improvement in stability, but at the expense of increased hopping activity, which increases delay and energy consumption. Adaptive FH, on the other hand, provides the highest level of consistency between all parameters: increased PDR, minimal delay, optimised energy consumption and stable switching frequency. Thus, the integrated analysis confirmed that the proposed adaptive frequency control method is the most effective solution for FANET swarm networks, ensuring communication reliability, energy efficiency and interference resistance at the same time.

Assessment of statistical reliability and reproducibility of FANET system modelling results

To confirm the reliability of the data obtained, a statistical analysis of the results of 30 simulation series covering three frequency control methods (fixed frequency, FHSS, and Adaptive FH) was performed. The normality of the distribution was verified using the Shapiro-Wilk criterion, and the homogeneity of the variances was checked using the Levene test. All samples showed normal distribution ($p > 0.05$) and homogeneity of variances ($p > 0.05$), which ensured the correct application of the parametric variance analysis ANOVA. Statistically significant inter-mode differences were recorded in cases where the p-value obtained during ANOVA was less than 0.05. The results of two-factor ANOVA with the factors "method type" and "distance between nodes" are presented in Table 4.

Table 4. Results of two-factor analysis of variance (ANOVA)

Indicator	F-value	p-value	Effect size (Cohen's d/d/Glass' Δ)
PDR (method type)	18.42	0.001	0.81/0.76
Delay (method type)	15.35	0.004	0.73/0.68
Energy (method type)	11.29	0.008	0.65/0.61
Switching Rate	9.84	0.01	0.59/0.55
PDR (distance)	7.92	0.013	-
Delay (distance)	5.64	0.021	-

Source: compiled by the authors based on statistical analysis of simulation results in MATLAB R2024a and Python 3.12

Table 4 shows that the effect of the frequency control method type is statistically significant for all key indicators: high F-values for PDR (18.42) and delay (15.35) indicate that the difference between the methods significantly exceeds the internal variability of the data. Cohen's d and Glass' Δ effect sizes in the range of 0.59-0.81 confirm the practical significance of these differences; the methods do indeed form different performance profiles, rather than merely demonstrating statistical deviations. The "distance" indicator also has an impact on PDR and delay ($p < 0.05$), but its effect is noticeably weaker, which is reflected in the absence of effect sizes for this factor. Thus, the decisive factor in changing the FANET parameters is the type of frequency control method, while the distance between nodes has an additional, but less pronounced, effect.

The results obtained comprehensively confirmed the effectiveness of the intelligent frequency management approach in FANET, based on a combination of fuzzy logic, adaptive frequency hopping, and autonomous routing between nodes. The proposed solution is based on a Mamdani-type fuzzy controller, which provides a dynamic response to changes in channel quality, noise level, residual energy, and node load. This created a self-learning frequency control system capable of maintaining communication stability without centralised control, a key property for FANET in real dynamic conditions.

A comparative analysis of three approaches – fixed frequency, FHSS, and Adaptive FH – revealed a fundamental difference in their operation. Fixed frequency provides a minimum number of switches, but demonstrates the lowest channel stability and significant packet loss in dynamic conditions. FHSS is more resistant to interference, but operates with an increased frequency of hopping and generates significantly higher latency with high node mobility. In this context, Adaptive FH provides an optimal combination of characteristics: it ensures a consistently higher packet delivery rate, lower latency, and lower power consumption, while reducing the number of hops compared to FHSS. Using fuzzy logic and responding to the channel status, the system adapts the frequency to specific conditions, which maintains communication quality even when the distance between nodes is increased. These properties confirm that Adaptive FH forms the most balanced FANET performance profile and has the best prerequisites for use in next-generation intelligent swarm networks.

Discussion

The results demonstrated a statistically significant advantage of adaptive frequency control using fuzzy logic over classical fixed frequency and FHSS schemes in all key indicators: packet delivery rate, delay, power consumption,

and switching frequency. This improvement confirmed the feasibility of combining fuzzy logic with adaptive frequency hopping, which was consistent with the findings of previous studies in the field of FANET intelligent routing. M. Aissa *et al.* (2025) determined that clustering based on fuzzy logic provided an 18% increase in network energy efficiency and topology stability through contextual adaptation to node dynamics. Compared to the results obtained, the study observed a similar reduction in energy consumption of 12-19%, indicating a similar trend towards energy stabilisation under the action of a fuzzy controller. S.M. Ahmed & A.S. Mohammed (2022) demonstrated the effectiveness of a fuzzy adaptive router in packet propagation in FANET, achieving a 22% increase in the delivery rate in noisy environments. Similarly, the proposed Adaptive FH method demonstrated a 25-30% improvement in PDR, confirming the versatility of fuzzy logic for compensating for losses in complex radio conditions.

P. Aimtongkham *et al*. (2024) demonstrated that a fuzzy system for channel overload control improved routing efficiency in low-power networks. A similar effect was observed in the results obtained, where adaptive frequency hopping minimised losses as the distance between nodes increased, ensuring a smooth reduction in PDR without sharp drops. M. Aalsalem (2023) demonstrated that neuro-fuzzy approaches significantly reduced transmission delay by predicting congestion, which correlated with the results obtained: the average end-to-end delay decreased to 43 ms compared to 78 ms in the baseline scenario. The effectiveness of fuzzy control mechanisms is consistent with the findings of R.I. Al-Essa & G.A. Al-Suhail (2023), demonstrating that adaptive beaconing based on fuzzy logic can reduce service transmissions and increase the stability of geographic routing. A similar effect is observed in the results obtained, where the switching frequency in Adaptive FH mode was reduced by almost half, indicating a more rational use of frequency resources. The approach proposed by S. Sugantha Priya & M. Mohanraj (2023) demonstrated that the use of fuzzy models in routing can increase the energy efficiency of nodes by approximately 17%. This conclusion fully correlates with the reduction in energy consumption of the adaptive frequency control scheme obtained in this work, confirming the versatility of fuzzy methods for optimising FANET. A systematic review by T.R. Beegum *et al.* (2023) emphasised that bio-inspired optimisation algorithms, particularly swarm and ant algorithms, are less sensitive to channel noise but often inferior to fuzzy models in terms of communication stability. The more uniform PDR growth in Adaptive FH observed during modelling is consistent with this relationship.

Fuzzy models have also shown advantages in transport networks. As noted by K.K. Jajala & R. Buduri (2024), combining ant optimisation with fuzzy control reduces channel congestion and minimises delays. The lowest end-to-end delay recorded in FANET using the Adaptive FH method confirms this trend and demonstrates the consistency of results across different types of infrastructure. The advantages of fuzzy logic in networks with dynamic topology were also noted by M. Sahare & P. Maheshwary (2023), recording an increase in throughput and stability of approximately 20%. The improvement in PDR recorded in the experiment is consistent with their conclusion and demonstrates the ability of the frequency controller to compensate for losses caused by node mobility. S. Badawi *et al.* (2025) emphasised adaptive frequency control in critical scenarios, stressing the relevance of intelligent frequency range change mechanisms for FANET in areas affected by radio interference. The results reflect this pattern: Adaptive FH maintained network performance even under conditions of intense narrowband jamming.

In FANET systems, the combination of trust-based control and fuzzy logic was efficient in improving communication reliability. The results of S. Alam *et al.* (2024) showed that route formation based on fuzzy trust assessment between nodes stabilised data transmission even under conditions of frequent topology changes. The results confirmed this trend, as adaptive frequency hopping maintained a high packet delivery rate (PDR = 0.93 – 0.95) in a dynamic environment. J. Kundu *et al.* (2025), proposing a socially oriented trust-based routing model, demonstrated that fuzzy trust weight calculation reduces the number of route failures. This result was consistent with experimental observations, where communication stability was maintained even with active node movement. The approach of A.T. Albu-Salih & H.A. Khudhair (2021) implemented the SDN architecture in ASR-FANET, reducing routing delay through dynamic channel reallocation, which was consistent with the reduction in average end-to-end delay to 43 ms achieved in this study. A similar role of flexible adaptation was confirmed by the results of K. Sun *et al.* (2025), applying a Kalman filter to predict channel status, ensuring transmission stability at variable node speeds. In the simulation, Adaptive FH achieved a similar effect of maintaining stable PDR with topology variations.

The combination of artificial intelligence with adaptive protocols proposed by P. Prabhakar *et al.* (2025) increased throughput and reduced latency by predicting traffic. A similar pattern was evident in the results obtained, where a fuzzy controller effectively managed frequency, ensuring a balance between speed and connection quality. The energy-saving concept presented by R. Sivaranjani *et al.* (2025) was also consistent with the identified trends: in the Adaptive FH system, energy consumption decreased by 12-19%, which is close to the figures recorded in the FLEATM model. The development by R.C. Karpagalakshmi *et al.* (2024), which combined bio-inspired optimisation with fuzzy zonal clustering, confirmed that hybrid approaches improve communication stability. A similar effect was observed here, with a 25-30% expansion of the communication working area compared to baseline methods. An analytical review by A.H. Wheeb *et al.* (2022) determined that decentralised models with predictive mechanisms respond more effectively to environmental changes. This is consistent

with the data obtained, where the fuzzy system provided stability without centralised coordination.

The study by A. Malhotra & S. Kaur (2022) emphasised that cognitive adaptive algorithms increase the efficiency of FANET through self-organisation. This principle was evident in the adaptive frequency hopping logic, which acted as a self-learning module. Analysis by E. Felemban (2021) proved that minimising delay is a key factor at high node speeds, and this effect was confirmed by the results of Adaptive FH maintaining the lowest Delay values. J. Vijitha Ananthi & P. Subha Hency Jose (2022) emphasised the need for multi-level integration of routing and frequency mechanisms to improve network stability. Within the experimental data obtained, such integration was implemented naturally through the interaction of a fuzzy controller and a frequency adapter, which ensured a coordinated reduction in delay and energy consumption. A review by M.J. Almansor *et al.* (2024) confirmed that multi-criteria control systems in FANET improve communication reliability; a similar effect was demonstrated by the developed model, where all parameters remained in a balanced ratio. The systematisation by S.A. Hasan *et al.* (2024) identified the need for protocols capable of adapting to unpredictable conditions, a pattern that was fully consistent with the stable behaviour of the Adaptive FH system in 30 simulation series.

Thus, the study expands the scientific discourse on the development of intelligent frequency management in FANET, empirically proving that the combination of fuzzy logic with adaptive frequency hopping provides a significant increase in the efficiency, stability, and energy efficiency of network interaction. The identified patterns confirmed that hybrid schemes prioritise contextual routing, and self-learning channel parameter control is used by FANET to maintain a high packet delivery rate, minimal transmission delays, and resistance to interference even in dynamic environmental conditions. Correlation of the results with previous studies showed that the key challenges of topology instability, limited energy resources, and noise immunity remain decisive for the efficiency of swarm networks. At the same time, the proposed approach has proven its ability to compensate for these factors through flexible fuzzy control that integrates channel quality, load level, and energy status data of nodes into a single decision-making model. Overall, the results confirm that the transition from static schemes to adaptive cognitive frequency management systems is setting a new standard for FANET networks, which are capable of autonomously analysing the environment, predicting changes in communication parameters, and optimising data transmission in real time. This approach creates prospects for the creation of intelligent, self-configuring network architectures for monitoring, coordinating and controlling unmanned systems in highly dynamic environments.

Conclusions

The results of the study confirmed that the implementation of intelligent frequency control in FANET based on fuzzy logic and adaptive frequency hopping provides a significant increase in the stability, energy efficiency, and speed of the communication system. Modelling conducted in MATLAB R2024a and Python 3.12 environments showed that the developed fuzzy controller is capable of autonomously making decisions about changing the operating channel, considering RSSI, SNR, E and Q. This ensures adaptability of the system to dynamic environmental conditions without centralised control and maintains high data transmission quality even in the presence of narrowband interference.

A comparative analysis of three methods – Fixed Frequency, FHSS and Adaptive FH – showed that the adaptive approach provides a 25-30% expansion of the communication working area, an increase in PDR to 0.93-0.95, a 12-19% reduction in energy consumption, and a reduction in average end-to-end transmission delay to 43 ms. The results of a two-factor analysis of variance (ANOVA) confirmed a statistically significant effect of the method type on all key parameters ($p < 0.05$), with the effect size for PDR (Cohen's $d = 0.81$, Glass' $\Delta = 0.76$) indicating a strong influence of the adaptive mechanism on communication quality. The data also show that the distance between nodes is a substantial factor, but intelligent frequency control partially compensates for its negative impact by maintaining network stability over longer distances. A generalised normalised analysis of PDR, Delay, Energy and Switching Rate indicators confirmed the balance of the proposed method, which combines high throughput with low power consumption and minimal switching frequency. This indicates a systemic compromise between speed, reliability, and spectrum efficiency, which is crucial for next-generation FANET networks.

Despite the positive results obtained, the study has certain limitations: the modelling was performed on a sample of five nodes and within a fixed topology, without incorporating complex obstacles or changes in flight altitude. Prospects for further research include scaling the model to larger swarm formations, integrating adaptive frequency control with machine learning-based routing protocols, and testing hybrid Fuzzy Logic + Reinforcement Learning schemes. This will make it possible to overcome the existing limitations of the model, enhance the generalisability of the obtained results, and establish a foundation for designing intelligent FANET networks suitable for deployment in real-world scenarios characterised by high dynamics and dense node interactions.

Acknowledgements

None.

Funding

The study was not funded.

Conflict of Interest

None.

References

[1] Aalsalem, M. (2023). An intelligent adaptive neuro-fuzzy for solving the multipath congestion in internet of things. *Journal of Information Systems Engineering and Management*, 8(4), article number 23845. doi: 10.55267/iadt.07.14044.

[2] Ahmed, S.M., & Mohammed, A.S. (2022). Fuzzy adaptive routing protocol for packet dissemination in FANET. *International Journal of Nonlinear Analysis and Applications*, 13(2), 1673-1683. doi: 10.22075/ijnaa.2022.27515.3633.

[3] Aimtongkham, P., Musikawan, P., Kongsorot, Y., & So-In, C. (2024). A novel congestion control scheme using fuzzy logic systems to enhance the path selection criteria in routing protocols for low-power and lossy networks on the internet of things. *SN Computer Science*, 5(5), article number 610. doi: 10.1007/s42979-024-02940-z.

[4] Aissa, M., Bouhdid, B., Bahri, M.A., Zakarya, M., & Mayyahi, K.A. (2025). Adaptive clustering in FANETs: A fuzzy logic approach for energy efficiency and network stability. *Telecommunication Systems*, 88(4), article number 125. doi: 10.1007/s11235-025-01356-1.

[5] Alam, S., Kundu, J., Ghosh, S., & Dey, A. (2024). Trusted fuzzy routing scheme in flying ad-hoc network. *Journal of Fuzzy Extension and Applications*, 5(1), 48-59. doi: 10.22105/jfea.2024.436052.1370.

[6] Albu-Salih, A.T., & Khudhair, H.A. (2021). ASR-FANET: An adaptive SDN-based routing framework for FANET. *International Journal of Electrical & Computer Engineering*, 11(5), 4403-4412. doi: 10.11591/ijece.v11i5.pp4403-4412.

[7] Al-Essa, R.I., & Al-Suhail, G.A. (2023). AFB-GPSR: Adaptive beaconing strategy based on fuzzy logic scheme for geographical routing in a mobile ad hoc network (MANET). *Computation*, 11(9), article number 174. doi: 10.3390/computation11090174.

[8] Almansor, M.J., Din, N.M., Baharuddin, M.Z., Ma, M., Alsayednoor, H.M., Al-Shareeda, M.A., & Al-asadi, A.J. (2024). Routing protocols strategies for flying Ad-Hoc network (FANET): Review, taxonomy, and open research issues. *Alexandria Engineering Journal*, 109, 553-577. doi: 10.1016/j.aej.2024.09.032.

[9] Alotaibi, J. (2025). FuzOptRoute: A fuzzy logic-integrated optimization-based energy-efficient cluster routing framework with edge computing for mobile communication networks. *The Journal of Supercomputing*, 81(11), article number 1186. doi: 10.1007/s11227-025-07673-1.

[10] Atheeq, C., Gulzar, Z., Al Reshan, M.S., Alshahrani, H., Sulaiman, A., & Shaikh, A. (2024). Securing UAV networks: A lightweight chaotic-frequency hopping approach to counter jamming attacks. *IEEE Access*, 12, 38685-38699. doi: 10.1109/ACCESS.2024.3375343.

[11] Badawi, S., Ahmad, N., Akmam, R., Mohamed, N., & Mohd, S. (2025). Routing protocols in FANET for disaster area networks: A review. *ASEAN Engineering Journal*, 15(3), 81-100. doi: 10.11113/aej.v15.22954.

[12] Beegum, T.R., Idris, M.Y., Ayub, M.N., & Shehadeh, H.A. (2023). Optimized routing of UAVs using bio-inspired algorithm in FANET: A systematic review. *IEEE Access*, 11, 15588-15622. doi: 10.1109/ACCESS.2023.3244067.

[13] Bhatia, T.K., Gilhotra, S., Bhandari, S.S., & Suden, R. (2024). Flying ad-hoc networks (FANETs): A review. *EAI Endorsed Transactions on Energy Web*, 11. doi: 10.4108/ew.5489.

[14] Bieliakov, R., & Fesenko, O. (2023). A model of intelligent resource management of class manet terrestrial communication network. *Information Technology and Society*, 3(9), 6-14. doi: 10.32689/maup.it.2023.3.1.

[15] Felemban, E. (2021). Evaluation of routing protocols and mobility in flying ad-hoc network. *International Journal of Advanced Computer Science and Applications*, 12(7), 643-650. doi: 10.14569/IJACSA.2021.0120773.

[16] Hasan, S.A., Mohammed, M.A., & Sulaiman, S.K. (2024). Flying ad-hoc networks (FANETs): Review of communications, challenges, applications, future direction and open research topics. *ITM Web of Conferences*, 64, article number 01002. doi: 10.1051/itmconf/20246401002.

[17] Hosseinzadeh, M., Husari, F.M., Yousefpoor, M.S., Lansky, J., & Min, H. (2024). A local filtering-based energy-aware routing scheme in flying ad hoc networks. *Scientific Reports*, 14(1), article number 17733. doi: 10.1038/s41598-024-68471-y.

[18] Jajala, K.K., & Buduri, R. (2024). Efficient and secure routing with UAV: GuidedPheromone update based on improved Ant colony optimization and fuzzy logic for congestion control in vehicular ad-hoc network. *International Journal of Information Technology*, 16(7), 4089-4110. doi: 10.1007/s41870-024-01978-9.

[19] Karpagalakshmi, R.C., Rani, D.L., Magendiran, N., & Manikandan, A. (2024). An energy-efficient bio-inspired mobility-aware cluster p-WOA algorithm for intelligent whale optimization and fuzzy-logic-based zonal clustering algorithm in FANET. *International Journal of Computational Intelligence Systems*, 17(1), article number 258. doi: 10.1007/s44196-024-00651-0.

[20] Khan, S., Khan, M.Z., Khan, P., Mehmood, G., Khan, A., & Fayaz, M. (2022). An ant-hocnet routing protocol based on optimized fuzzy logic for swarm of UAVs in FANET. *Wireless Communications and Mobile Computing*, 2022(1), article number 6783777. doi: 10.1155/2022/6783777.

[21] Kundu, J., Alam, S., & Dey, A. (2025). Fuzzy logic based social trust computation scheme in Flying Ad-hoc network. *Journal of Fuzzy Extension and Applications*, 6(1), 59-70. doi: 10.22105/jfea.2024.445025.1385.

[22] Malhotra, A., & Kaur, S. (2022). A comprehensive review on recent advancements in routing protocols for flying ad hoc networks. *Transactions on Emerging Telecommunications Technologies*, 33(3), article number e3688. doi: 10.1002/ett.3688.

[23] Mamdani, E.H., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, 7(1), 1-13. doi: 10.1016/S0020-7373(75)80002-2.

[24] Prabhakar, P., Yokesh, V., Aruchamy, P., & Nanthakumar, S. (2025). Artificial intelligence-enabled fully echoed Q-routing and adaptive directional medium access control protocol for flying ad-hoc networks. *International Journal of Communication Systems*, 38(4), article number e6138. doi: 10.1002/dac.6138.

[25] Prakash, M., Neelakandan, S., & Kim, B.H. (2024). Reinforcement learning-based multidimensional perception and energy awareness optimized link state routing for flying ad-hoc networks. *Mobile Networks and Applications*, 29(2), 315-333. doi: 10.1007/s11036-023-02255-y.

[26] Rahmani, A.M., Ali, S., Yousefpoor, E., Yousefpoor, M.S., Javaheri, D., Lalbakhsh, P., Ahmed, O.H., Hosseinzadeh, M., & Lee, S.W. (2022). OLSR+: A new routing method based on fuzzy logic in flying ad-hoc networks (FANETs). *Vehicular Communications*, 36, article number 100489. doi: 10.1016/j.vehcom.2022.100489.

[27] Sahare, M., & Maheshwary, P. (2023). The congestion control and performance improvement by fuzzy techniques in FANET with IoT: A survey. In *Proceedings of the international conference on ICT in business industry & government* (pp. 1-8). Indore: IEEE. doi: 10.1109/ICTBIG59752.2023.10455986.

[28] Semendiai, S. (2023). The use of cognitive radio technology to improve the efficiency of wireless data transmission systems in the conditions of active use of electronic warfare. *Cybersecurity: Education, Science, Technique*, 4(20), 220-229. doi: 10.28925/2663-4023.2023.20.220229.

[29] Sivaranjani, R., Shankar, R., & Duraisamy, S. (2025). FLEATM: Fuzzy logic-based energy-aware trust based routing in MANETs. In S. Rajagopal, K. Popat, D. Meva, S. Bajeja & P. Mudholkar (Eds.), *International conference on advancements in smart computing and information security* (pp. 144-159). Cham: Springer. doi: 10.1007/978-3-031-86296-0_12.

[30] Sugantha Priya, S., & Mohanraj, M. (2023). An energy-efficient clustering and fuzzy-based path selection for flying ad-hoc networks. *International Journal of Computational Intelligence and Applications*, 22(1), article number 2341003. doi: 10.1142/S1469026823410031.

[31] Sun, K., Liu, M., Yin, C., & Wang, Q. (2025). Adaptive extended Kalman prediction-based SDN-FANET segmented hybrid routing scheme. *Sensors*, 25(5), article number 1417. doi: 10.3390/s25051417.

[32] Valuyskyi, S., & Ponomarenko, O. (2025). Analysis of the architecture, technologies and future challenges of flying ad hoc networks. In *Proceedings of the international scientific conference "Modern challenges in telecommunications"* (pp. 345-347). Kyiv: National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

[33] Vijitha Ananthi, J., & Subha Hency Jose, P. (2022). A review on various routing protocol designing features for flying ad hoc networks. In S. Shakya, R. Bestak, R. Palanisamy & K. Kamel (Eds.), *Mobile computing and sustainable informatics: Proceedings of ICMCSI 2021* (pp. 315-325). Singapore: Springer. doi: 10.1007/978-981-16-1866-6_23.

[34] Wheeb, A.H., Nordin, R., Samah, A.A., Alsharif, M.H., & Khan, M.A. (2022). Topology-based routing protocols and mobility models for flying ad hoc networks: A contemporary review and future research directions. *Drones*, 6(1), article number 9. doi: 10.3390/drones6010009.

Інтелектуальне управління частотою в FANET: нечітка логіка/маршрутизація і адаптивний frequency hopping

Роман Зайвий

Аспірант
Національний університет «Львівська політехніка»
79000, вул. Степана Бандери, 12, м. Львів, Україна
https://orcid.org/0009-0003-4096-4111

Володимир Павлиш

Кандидат технічних наук, професор
Національний університет «Львівська політехніка»
79000, вул. Степана Бандери, 12, м. Львів, Україна
https://orcid.org/0009-0004-3996-5923

Анотація. Метою дослідження було експериментальне оцінювання ефективності інтелектуального управління частотою у роєвих мережах безпілотних літальних апаратів (БпЛА) із використанням нечіткої логіки та адаптивного перестрибування частоти. Об'єктом аналізу виступили три методи частотного керування – фіксована частота, класичне перестрибування частоти та запропонований адаптивний метод, який поєднує нечітке логічне прийняття рішень із контекстно-залежною маршрутизацією. Дослідження проводилося у симуляційному середовищі MATLAB R2024a та Python 3.12 на моделі з п'яти БпЛА, що рухалися у межах області розміром 1000×1000 м із урахуванням змін топології, рівня сигналу, співвідношення сигнал/шум та енергетичних характеристик вузлів. Отримані результати показали, що розроблений адаптивний метод забезпечує найвищу ефективність зв'язку серед досліджених підходів. Коефіцієнт доставки пакетів утримувався на рівні 0,93–0,95 навіть за наявності вузькосмугових перешкод, що на 25–30 % перевищує показники базових методів. Середня наскрізна затримка передачі зменшилася до 43 мс проти 61 мс у класичній схемі перестрибування частоти та 78 мс у фіксованому режимі. Енергоспоживання знизилося на 12–19 %, а середня частота перемикань скоротилася удвічі (≈ 2 рази/с проти 4,2 раза/с у класичному режимі), що свідчить про оптимізацію роботи контролера. Статистичний аналіз підтвердив значущий вплив типу методу на всі основні показники ефективності зв'язку ($p < 0,05$), що засвідчує достовірність отриманих результатів і відтворюваність системи у серії симуляційних експериментів. Запропонований підхід забезпечує автономну оптимізацію маршрутів передавання даних і підтримання стабільного каналу зв'язку навіть у динамічних середовищах, що відкриває перспективи для розроблення нового покоління інтелектуальних мереж БпЛА, орієнтованих на завдання моніторингу, розвідки та координації у реальному часі. Результати дослідження можуть бути використані розробниками безпілотних систем, інженерами зв'язку та фахівцями з мережевих технологій для створення більш стійких до завад, енергоефективних і самонавчальних систем зв'язку

Ключові слова: безпілотні літальні апарати; роєві мережі; радіомодуль; енергоефективність; коефіцієнт доставки пакетів; мобільність вузлів

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 28.08.2025 Revised: 14.11.2025 Accepted: 23.12.2025

UDC 004.85 DOI: 10.31649/vitce/3.2025.54

Active self-learning for object detection in an imbalanced data environment: The TAAST approach

Dmytro Ivanov*

Posrgraduate Student
Zhytomyr Polytechnic State University
10005, 103 Chudnivska Str., Zhytomyr, Ukraine
https://orcid.org/0000-0002-7386-4497

Abstract. In the context of the growing development and application of computer vision, there is a growing need to reduce the cost of manual data markup, especially in tasks of detecting rare objects in conditions of long-tailed class distribution. The purpose of the study was to improve the efficiency of identifying rare image categories by improving the active self-learning strategy. The study used the Tail-Aware Active Self-Training approach, which was based on strategic selection of frames, considering the entropy of uncertainty, class rarity, and semantic diversity in the feature space of the Contrastive Language-Image Pretraining model, followed by the use of pseudo-markup using the You Only Look Once detector, version 8. As a result of experiments on Large Vocabulary Instance Segmentation datasets, version 1.0, and nuImages-imbalanced, the proposed strategy provided an increase in AP_rare accuracy by 6.3-6.4 percentage points compared to the basic Random and Uncertainty Sampling approaches. The overall accuracy of the model did not decrease, but increased to 36.0-43.2% mAP, depending on the dataset. The markup efficiency indicator reached 42-43%, which was 9-10 points higher than competitive strategies. The results of the experiment were statistically reliable, since the confidence intervals for the AP_rare accuracy metric in the case of using the Tail-Aware Active Self-Training method do not overlap with the intervals for the basic random and Uncertainty-only strategies. This indicated that the advantage of this method was not random, but was confirmed with high probability. Consequently, the results obtained demonstrated the reliability and stability of the proposed approach. It was demonstrated that after two active iterations, the model reached a performance plateau, which significantly reduced computational costs. The practical significance of the study lies in creating an effective tool for automated deployment of computer vision models in conditions of a limited markup budget

Keywords: machine learning; semantic clustering; pseudoanotation; entropy sampling; class balancing; computer vision; markup optimisation

Introduction

Computer vision systems are rapidly developing, and advanced object detection models demonstrate high accuracy on balanced data sets. However, in real-world problems, images often have a long-tailed distribution: most objects belong to categories with low representation in the sample. Under such conditions, models lose their ability to effectively generalise to rare classes, which is critical for applications in biomonitoring, autonomous driving, safety, etc. This problem is becoming particularly relevant due to the growing need for automated data processing in high-risk or hard-to-reach environments, where markup for a large number of images is extremely resource-intensive.

Recent studies confirm that the main reason for the low efficiency of object detection models in rare classes is a pronounced class imbalance in image sets. In particular, in LVIS (Large Vocabulary Instance Segmentation), nuImages, and iNaturalist, most classes have less than 10 examples, which reduces the quality of recognition and is masked by the global mean Average Precision (mAP) metric. As noted by Y. Li *et al.* (2020), the distribution of objects in the LVIS v1.0 Set obeys Zipf's law: only a third of classes have more than 100 examples, and more than 28% have less than 10. A similar situation was described by H. Caesar *et al.* (2020) for the nuImages dataset designed for realistic autonomous driving scenes – more than half of the classes occur less than 10

Suggested Citation:
Ivanov, D. (2025). Active self-learning for object detection in an imbalanced data environment: The TAAST approach. *Information Technologies and Computer Engineering*, 22(3), 54-64. doi: 10.31649/vitce/3.2025.54

*Corresponding author

times. In an analysis of the iNaturalist dataset for biodiversity tasks (De Alvis & Seneviratne, 2024), the ratio between common and rare classes exceeds 1:500, which reduces the accuracy of the latter by almost 20 percentage points. Such results show that a global metric such as mAP can mask critically poor quality for underrepresented categories.

To mitigate the impact of the imbalance in long-tailed object detection problems, the researchers proposed a number of modifications to the loss functions and classification layers. The Balanced Group Softmax method (Li *et al.*, 2020) implements group normalisation of logits in accordance with the frequency of classes, which provides an increase in ap_rare accuracy by 4-5 percentage points. Within the unified Balanced Classification scheme (Qi *et al.*, 2023) an approach to loss weighting is generalised, allowing adaptation to different degrees of class representation. The method developed by B. Li *et al.* (2022) with multimodal learning using pseudo labels at the image level also demonstrates improved accuracy on low-frequency classes. C.-L. Duan *et al.* (2024) presented the DRCL-2 approach, which combines contrast training with the reconstruction task and helps to further increase AP_rare by 5 percentage points.

However, the above methods are model-oriented, i.e., they involve complete markup of data sets, which does not solve the problem of high time and resource costs. In this context, active learning is promising, where the model requests annotations only for the most informative samples. Y. Gal & Z. Ghahramani (2016) proposed Entropy Sampling, a strategy for selecting images with maximum forecast entropy. The CoreSet Sampling approach, presented by O. Sener & S. Savarese (2018), provides sample diversity through clustering in feature space. J. Wu *et al.* (2022) integrated these ideas into the Entropy-AL + Progressive Diversity method, which increased mAP by 3 percentage points within a fixed budget on the COCO (Common Objects in Context) set. As indicated by K. Sohn *et al.* (2020) and M. Xu *et al.* (2021), most active strategies do not consider class rarity – the annotation budget is often spent on already well-represented categories, while rare classes are ignored.

Thus, existing approaches are either aimed at improving accuracy in rare classes without reducing markup costs, or optimise the annotation budget without considering class rarity. The lack of methods that combine both approaches – focusing on rare classes and saving manual markup – creates a noticeable gap in current research. The purpose of this study was to develop and empirically test an active self-learning strategy for the object detection problem focused on rare classes under long-tailed distribution conditions. To achieve this goal, the following is proposed: sampling by uncertainty, weighted by class frequency; clustering to ensure diversity in the space of semantic features obtained using CLIP (Contrastive Language-Image Pre-training); pseudo-labelling with a confidence threshold of 0.8. The study's hypothesis was that this approach would reduce the amount of manual annotation while maintaining or even improving accuracy in rare classes.

Materials and Methods

The developed Tail-Aware Active Self-Training (TAAST) approach is presented as a formalised technique that covers the full cycle of active self-learning training of the object detection model – from using the initial seed set and pseudo-marking to selecting informative examples and further training of the detector. Formalisation in the form of an algorithm and optimisation problem guarantees reproducibility of experiments and provides a reasonable assessment of the effectiveness of the strategy. Formalisation of the cycle of the active-self-learning process of training a model for the problem of object detection in conditions of limited access to annotated data was carried out according to a typical practical scenario, in which:

- a large array of unanotated images U is available (for example, from cameras of driver assistance systems – Advanced Driver-Assistance Systems (ADAS); aerial photos from drones; log files of multi-season monitoring, etc.);
- a small initial set of annotated data L is available, covering approximately 10% of the total volume – such a "seed set" is usually formed as part of the pilot stage;
- fixed budget B of frames allocated for each iteration of active learning;
- total number of active iterations is denoted as T.

To maximise the accuracy of the model in rare classes with the minimum possible volume of new annotated examples, it was proposed to integrate active learning with a self-learning approach, where the model used its own predictions to expand the learning set. The target metric for evaluating the effectiveness of the proposed method was the average accuracy value calculated separately for rare classes (mAP_rare). This helped to focus attention on exactly the subset of objects that is traditionally most vulnerable to imbalances and lack of training examples.

The YOLOv8-s model (version 8, small configuration) was chosen as the basis for the system, which demonstrates a sufficient level of accuracy (~38% mAP) on the COCO set with a significantly lower computational load compared to larger variants (Jocher *et al.*, 2023). A special feature of this architecture is the use of an anchor-free detection head, which does not require preliminary determination of object sizes and better summarises examples that rarely occur in the training set, in particular, on the "long tail" of the distribution (Tian *et al.*, 2019). Based on initialisation with weights previously trained in COCO, the model is already able to generate fairly accurate pseudo-labels in the first active cycle without additional configuration. To calculate the semantic similarity of scenes, the CLIP model with the VIT-L/14 (Vision Transformer) architecture was used, which was trained on paired text – image examples and can encode plot features in the form of compact 512-dimensional vectors (Radford *et al.*, 2021). Clustering of these vectors was performed using an algorithm k-means++, which provided fast and stable splitting of a large number of vectors into groups due to improved centre initialisation (Johnson *et al.*, 2021). After filtering by value and diversity, the frames were grouped into B clusters (by the number of

available signatures), and the representative frame closest to the centre was selected from each cluster. The model was further trained on a combined set that contained initial human annotations, new marked frames, and pseudo-annotations with high confidence (⩾ 80). Optimisation was performed using stochastic gradient descent (SGD) with a momentum of 0.937 and a regularisation coefficient (weight decay) of $5 \cdot 10^{-45}$, one of the most effective optimisation methods in machine learning problems (Bottou, 2012). The warm-up (initial phase) lasted 300 epochs with a linear decrease in the learning rate from 0.01 to 0.001, while each subsequent active cycle covered only 30 epochs, which helped to quickly adapt the model to new data without overtraining.

All experimental studies were conducted on two commonly used datasets: LVIS v1.0 and nuImages-imbalanced (an imbalanced version of the nuScenes subset). For each of them, a controlled division scheme was applied into training and validation parts. In particular, 10% of the data was randomly selected from the initial training sample to form the initial body with manual marking, which is further designated as L_0 (seed-dataset). The remaining 90% of the training images formed a U pool that simulated a realistic situation of incomplete markup when starting a new data collection project. This distribution allows simulating the conditions of a limited human resource at the beginning of active training.

In this paper, a class was considered rare if it was found in less than 10 examples in the initial training set, which meets the LVIS-taxonomy criteria (Li *et al.*, 2022). To objectively evaluate the performance of the model, validation subsets were used, which remained fixed throughout all stages of the experiment. Evaluation of the test sample was performed only once – after all active cycles were completed, to avoid information leakage and re-evaluation of the results.

All active learning strategies in the study were implemented in three consecutive iterations ($T = 3$), which was chosen empirically: in two cycles, the potential of rare classes was not yet exhausted, while after the fourth cycle, the increase in the average accuracy metric for rare classes became less than 0.3 percentage points. In each cycle, the model generated a pool of pseudo-annotated examples P, adding to it all the provided objects for which the model confidence level exceeded the threshold of 0.8. Next, the top-20% filter was used for the integral significance score (x), which allowed excluding examples with too low a value and reduce duplication of head scenes. Semantic clustering was performed in this upper quintile subset, and one representative frame was selected from each cluster, for a total of $B = 256$ images per cycle. After each active cycle, the model was further trained on the combined set with $L \cup Q \cup P$ for 30 epochs using stochastic gradient descent and cosine reduction of the learning rate (from 0.01 to 0.001), in accordance with the recommendations for YOLOv8 (Jocher *et al.*, 2023). The same set of hyperparameters and procedure was used for all experimental strategies, which ensured the purity of comparison and made implicit reconfiguration impossible. As part of the experimental study, a consistent comparison of Random → Uncertainty-only → TAAST strategies was performed.

Results and Discussion

Stages of implementing the Tail-Aware Active Self-Training strategy

A detailed description of the sequence of actions within one active cycle of active self-learning of the object detection model ensures the reproducibility of the experiment. In addition, it helps to clearly understand the proposed method and evaluate its effectiveness. Below is a step-by-step scheme for implementing the TAAST strategy, where each step reflects the logic of the transition from automatic pseudo-markup generation to an optimisation training goal.

Step 1. Pseudo-markup based on confidence forecasts

The first stage of the proposed active self-learning strategy was to automatically expand the training set using the most reliable model predictions. This approach helped to reduce the amount of manual marking, while maintaining the quality of the training signal. For each object detected by the model in the unsigned image pool U, the level of trust in the object's belonging to a certain class was calculated. The object was moved to a set of pseudo-labels P, if its highest predicted probability exceeded a pre-determined confidence threshold. This was formalised by the following equation (proper formulation):

$$p_{max} = \max_k p_k \geq \tau, \tau = 0{,}8, \quad (1)$$

where p_k – probability that the detected object belongs to k-th class; $\max_k$ – maximum probability among all classes, i.e., the model's confidence in the most probable hypothesis; τ – confidence threshold is set at 0.8 (or 80%).

Selection of a threshold value $\tau = 0{,}8$ was based on previous experiments, where it was found that this level of trust provides an optimal compromise between the number of examples added and the noise level in pseudo-marking. Too low values of τ lead to a large number of false examples, while too high ones reduce the effectiveness of increasing the training set due to a limited number of confident forecasts.

Step 2. Assessment of frame value by uncertainty

After the most reliable predictions of the model are transferred to the pseudo – markup set, the next step is to evaluate the value of the remaining images from the unlabelled pool. It is necessary to select those examples that, when labelled manually, will bring the greatest increase in accuracy. The main criteria for such an assessment are the uncertainty of the model in relation to a particular frame, the presence of rare classes, and its diversity in the context of the entire sample. To quantify the uncertainty of the model with respect to the image , the sum of entropies is used for all objects detected in this image. In particular, the calculation is performed using the equation (adapted from C.-L. Duan *et al.* (2024)):

$$E(x) = \sum_{b \in \hat{y}(x)} \left(-\sum_{k=1}^{C} p_{k,b} \log p_{k,b} \right), \quad (2)$$

where $\hat{y}(x)$ – set of all provided objects (boxes) in the image x obtained from the model; b – separate box, i.e., a rectangular area that corresponds to the detected object; $p_{k,b}$ – probability that box b belongs to k-th class; C – total number of classes.

Entropy, as a measure of uncertainty, increases when the probability distribution is "flat", meaning that the model does not have a clear advantage in favour of any class. Accordingly, a high value of $E(x)$ indicates the complexity of the image for the model and the feasibility of marking it manually. This approach allows focusing limited resources on those images that can significantly improve the training of the model in the active cycle.

Definition of a candidate Image class in active learning

After estimating the overall uncertainty, it is necessary to determine the class of objects for which the model shows the greatest confusion in a particular image. To do this, all the predicted objects (frames, or boxes) on the frame are analysed, and the one in which the model has the lowest overall confidence is selected – that is, even the highest probability of belonging to any class is low. This allows identifying the "weakest point" for the model in a given image and the corresponding class as a candidate for improvement using manual annotation. Formally, this process is defined as follows (proper wording):

$$b^* = \arg \min_{b \in \hat{y}(x)} \left(\max_k p_{k,b} \right), \quad (3)$$

where x – image being analysed; $\hat{y}(x)$ – set of all provided objects (boxes) in the image x; $b \in \hat{y}(x)$ – specific frame within this image; $p_{k,b}$ – probability that the frame b belongs to the class k; $\max_k p_{k,b}$ – highest probability, which reflects the model's confidence level in its forecast for the frame b.

Thus, b^* indicates the frame for which the model is least confident, even in terms of its strongest prediction. Further, for this frame, the so-called candidate image class is defined – the class that the model still considers most likely for the frame b^* (actual wording):

$$c(x) = \arg \max_k p_{k,b^*}. \quad (4)$$

Class $c(x)$ is considered a representative of the category that the model confuses most in the image x. If the detected candidate class belongs to rare categories, the image gets a higher priority for subsequent manual markup as part of active learning. This allows effectively using a limited annotation resource, focusing it on examples that help to improve the accuracy of the model on rare classes.

Evaluation of the current representation of a class to determine its rarity

The next step is to evaluate how well the candidate class is represented $c(x)$, which caused the most uncertainty in the model in the current training set. To do this, the number of available examples of this class is calculated considering both manually annotated and pseudo-labelled samples. Evaluation is performed using the following expression (author's wording):

$$n_{c(x)} = |L_c(x)| + |P_{c(x)}|, \quad (5)$$

where $c(x)$ – class of object that the model considers most likely in the most unreliable area of the image x; $L_c(x)$ – subset of manually marked-up class images $c(x)$, included in the training set L; $P_{c(x)}$ – subset of class images $c(x)$ that were automatically added as pseudo-markings to the set P; $|\cdot|$ – operator that defines the number of elements in a set.

This equation allows quantifying the "saturation" of an individual class in the current data set. Low value of $n_{c(x)}$ indicates that the corresponding class is still rare, and new examples involving it may be of high value in the context of active selection. A high value means that the class is already sufficiently represented, and additional marking of frames with its presence is less of a priority.

Calculation of the frequency weight for rare classes in the sample

In order to give preference to images that contain rare categories when ranking examples, the number of available examples of the class $n_{c(x)}$ is converted to a weighting factor that is inversely dependent on the frequency of this class. This weight is determined by the following equation (author's wording):

$$\omega_{c(x)} = \frac{1}{\log(n_{c(x)} + \beta)} \quad (6)$$

where $\omega_{c(x)}$ – weighting factor for the class clas $c(x)s$, which is used later to prioritise the frame; $n_{c(x)}$ – total number of class images $c(x)$ available in the training set (both manually annotated and obtained as a result of pseudo-markup), calculated according to equation (5); β – positive constant that guarantees the certainty of a logarithmic function even when the number of examples of the class is zero; in this paper, the value is assumed to be $\beta = 1$.

This equation allows compensating for bias in favour of frequently presented classes. Due to logarithmic smoothing of values, the weighting factor $\omega_{c(x)}$ increases for classes with few examples and decreases for well-represented classes. Thus, even on a limited budget, active training with high priority selects those frames that can improve the accuracy of the model in poorly represented categories.

Calculation of the integral value of a frame for further example selection

To make an effective comparison between all images left without annotations, it is proposed to combine two previously calculated characteristics – the uncertainty of the model with respect to the image and the frequency weight of the associated class – into a single integral indicator. This indicator is determined by the equation (author's wording):

$$\varphi(x) = \omega_{c(x)} \cdot E(x), \quad (7)$$

where $\varphi(x)$ – total (integral) value of the image , which reflects its importance for further markup; $\omega_{c(x)}$ – class

frequency weight $c(x)$, to which the model gives the highest (but not yet certain enough) probability; this coefficient is calculated by equation (6) and is higher for rare classes; $E(x)$ – total entropy of all detected objects in the image x, which characterises the degree of uncertainty of the model.

This equation allows combining the semantic complexity of the frame (due to entropy) with information about the relevance of the class (due to frequency weight), which makes it an effective criterion for selecting examples for manual annotation. Therefore, all unmarked images are sorted by value $\varphi(x)$, and the frames with the highest values are selected for manual markup. This approach allows allocating a limited annotation budget to the most valuable examples for training the model.

Step 3. Selection of different frames based on semantic diversity

After each unsigned image , an integral value was assigned $\varphi(x)$ (equation 7), it is necessary to select the frames that will most contribute to improving the accuracy of the model. These are images that have a high potential for information content and help to cover rare categories of objects. Since manual markup has a limited budget, it is marked as B (number of images that can be annotated at each iteration), it is important not only to identify the most valuable samples from the standpoint of integral metrics, but also to ensure their diversity.

The selection process is implemented in two stages:

1. Filtering by value – all images are sorted in descending order by function value $\varphi(x)$, after which a preliminary pool of candidates is formed, which includes $B' > B$ examples with the highest scores. Value B' is set empirically (for example, within 2-3 × of B), to provide sufficient space for the next step – diversification.

2. Ensuring diversity – to avoid excessive repetition of similar scenes or classes among the selected samples, a clustering mechanism is applied in the feature space. This study utilised embeddings obtained using a pre-trained CLIP model. Images from the previous pool are grouped using an algorithm of k-means, and the closest representative to the centroid is selected from each cluster. This forms the final set of B-images that will be submitted for manual marking.

This approach allows combining information content (high values $\varphi(x)$ with a variety of samples. This is crucial to ensure generalisability of the model and avoid over-training it on too uniform examples. In addition, it makes optimal use of the limited resources of human annotation within the active self-learning cycle.

Pre-filtering by integral value

At the first stage of selecting images for manual marking, a preliminary cut-off of unpromising frames was performed. This allows focusing computing resources and human attention on the most informative examples and thereby increasing the effectiveness of active learning. To form the previous set of priority examples, a subset is defined S_φ:

$$S_\varphi = \{x \in U \setminus P \mid \varphi(x) \in \text{топ} - 20\%\}, \quad (8)$$

where U – multiple of all unassigned images; P – subset of images that are already included in the pseudo-markup set; $E(x)$ – total entropy of all detected objects (frames) in the image x, which characterises the degree of uncertainty of the model; $x \in U \setminus P$ – images that remain unsigned and were not automatically annotated; $\varphi(x)$ – integral value of the frame , calculated by equation 7; S_φ – subset of the highest priority images included in the top 20% by value $\varphi(x)$.

This procedure generates many examples that are potentially most useful for manual annotation, since they combine high model uncertainty and belonging to rare classes. This approach allows reducing computational costs and using the annotation budget more efficiently, avoiding the cost of insignificant snapshots. Validity of choosing a threshold value $p = 20\%$ is confirmed by the results of previous research in the field of active learning for object detection tasks, in particular, in the papers Entropy + Progressive Diversity (Wu *et al.*, 2022) and SoftTeacher-AL (Xu *et al.*, 2021), where it is recommended to use filtering in the range of 15-25% of the most valuable examples.

Transition to the space of semantic features.

At this stage, each image x from the set S_φ is converted to a compact numerical representation – a feature vector that preserves the semantic content of the image. The purpose of this transformation is to provide a space structure in which similar frames are located close to each other, and dissimilar frames are located at a greater distance. This allows effectively applying grouping methods, in particular clustering:

$$z(x) = \frac{f_{CLIP}(x)}{\|f_{CLIP}(x)\|_2} \in R^{512}, \quad (9)$$

where x – images from the set S_φ, which is pre-selected as a set of valuable frames (equation 8); $f_{CLIP}(x)$ – 512-dimensional feature vector obtained using a pre-trained CLIP model (Radford *et al.*, 2021), which displays the content of the image; $\|f_{CLIP}(x)\|_2$ – L2-norm (Euclidean length) of the feature vector; $z(x) \in R^{512}$ – normalised vector in the 512-dimensional feature space, which is used as input for further clustering, for example by the k-means method.

This mapping of semantic features into the space allows each image to match a unified numerical representation, or conditionally – its "digital DNA". This makes it easier to analyse similarities between scenes and avoids duplication when selecting images for manual markup. The use of normalised vectors ensures that clustering will be based solely on directions in the feature space, and not on absolute values of components, which is especially important when using cosine distance-based metrics.

K-mean clustering: Detection of typical scenes

After all selected images have been converted to normalised feature vectors using the CLIP model, each image x gets a vector representation $z(x) \in R^{512}$, which is placed in a common semantic space. In this space, scenes that are similar in content have close coordinates. The next step is to divide the set of these vectors into B clusters – this is exactly how many examples are planned to be submitted for manual annotation in the current active learning cycle. The

classical algorithm is used for this purpose *k*-means with improved centroid initialisation using the *k*-means++ method (Radford *et al.*, 2021). Mathematically, the problem is formulated as minimising the total square of the Euclidean distance between vectors and cluster centres (adapted from A. Radford *et al*. (2021)):

$$\min_{\mu_1,\ldots,\mu_k} \sum_{j=1}^{k} \sum_{x \in C_j} \left|\left| z(x) - \mu_j \right|\right|_2^2, k = B, \quad (10)$$

where B – number of clusters (equal to the frame budget for markup); $z(x) \in R^{512}$ – normalised feature vector obtained from the clip model for the image x; C_j – multiple images assigned to j-th cluster; $\mu_j \in R^{512}$ – centre of j-th cluster, calculated as the arithmetic mean of vectors in C_j; $||z(x) - \mu_j||_2^2$ – square of the Euclidean distance between the image vector and the centre of the cluster. This approach allows creating a representative sample of frames that are very diverse in content, which reduces redundancy and increases the efficiency of manual marking.

Creating an active markup set.

After clustering semantic vectors by the *k*-means method, each cluster C_j where $j = 1,\ldots, B$ represents a group of frames that are similar in content. Next, an active set for manual markup is generated: one of the most representative examples is selected from each cluster. This approach allows ensuring maximum coverage of the content space with a fixed budget for markup. Calculating the active sub-set Q is performed according to the following equation:

$$Q = \left\{ x_j^* | x_j^* = arg \min_{x \in C_j} ||\, z(x) - \mu_j ||_2, j = 1, \ldots, B \right\}, \quad (11)$$

where C_j – j-th cluster formed as a result of the algorithm of *k*-means; all frames in the middle C_j have similar semantic features; $\mu_j \in R^{512}$ – centre of j-th cluster, calculated as the average value of vectors $z(x)$ for all $x \in C_j$; $z(x) \in R^{512}$ – normalised feature vector obtained from the CLIP model; $||\, z(x) - \mu_j ||_2$ – Euclidean distance between the frame feature vector x and the centre of the corresponding cluster; $arg \min_{x \in C_j}$ – operator that returns the frame with the smallest distance to the centre of the cluster, i.e., the most typical frame within the cluster C_j; x_j^* – frame that best represents the cluster C_j; Q – subset of B images, each of which is selected from a different cluster. Thus, the constructed set Q ensures that each markup frame represents a unique type of scene. This can significantly increase the efficiency of spending limited human resources, reducing redundancy and helping to speed up the process of self-learning the model for object detection.

Step 4. Updating and retraining the model on the combined data set

After a set of images Q marked up by experts on the previous one, manually adds annotated data, and confident pseudo-markings are collected, the stage of additional configuration of the model on the combined sample is performed. This section presents three key equations that formalise the structure of the new training set and the process of optimising the model weights. Updating of the manual dial:

$$L^{new} = L \cup Q, \quad (12)$$

where L – multiple images that were previously marked up manually; Q – multiple frames that were annotated by experts in the current iteration; L^{new} – updated manually marked-up set. Repetitions (duplicates) are automatically deleted, so each image is presented only once.

Creation of a complete training set:

$$T^{train} = L^{new} \cup P, \quad (13)$$

where P – set of pseudo-markings obtained on the basis of confidence forecasts of the model with a confidence threshold of at least 0.8; T^{train} – combined training sample that includes both human and automatically generated markings.

Model optimisation procedure (adapted from A. Radford *et al*. (2021)):

$$\theta^{(t+1)} = \theta^t - \eta \nabla_\theta L(T^{train}; \theta^t), t = 0,\ldots, e-1, \quad (14)$$

where θ^t – model parameters at the beginning of epoch t; η – learning rate, which gradually decreases from 0.01 to 0.001 according to the cosine attenuation graph; $L(T^{train}; \theta^t)$ – loss function that combines classification and regression components specific to the YOLO architecture (Ali & Zhang, 2024); $e = 30$ – number of epochs of additional training.

After completing 30 epochs, the model updates its scales to reflect new patterns, while maintaining previous knowledge. If the number of active iterations T did not reach the specified maximum, the process returns to the beginning of the cycle, in particular, to the pseudo-marking stage, which ensures the integration of active learning with self-learning.

Step 5. Statement of the optimisation goal of the active cycle for rare classes

After a detailed review of the stages of pseudo-markup, adjusted selection of examples and additional training of the model, it is necessary to formalise the target function of active learning and the corresponding restrictions. This section defines what exactly needs to be optimised and what resources contain the best strategy. The optimisation goal is to maximise recognition accuracy for rare classes after completing the entire sequence of active loops. Formally, the objective function is written as follows:

$$\max_S AP_{rare}(M_\theta^{(T)}), \quad (15)$$

where $M_\theta^{(T)}$ – detector with parameters θ after completion of T iterations of active learning; AP_{rare} – mean accuracy for rare classes only according to the LVIS v1.0 taxonomy; S – example selection strategy that considers the uncertainty, rarity, and variety of scenes in this case.

This goal is consistent with approaches in active learning, in particular, with the wording by B. Settles (2009), however, with a particular focus on rare classes. The limit on the manual markup budget is set by the inequality:

$$|L| \leq M_0 + BT, \quad (16)$$

where $|L|$ – total number of examples that were marked up manually after all cycles were completed; M_0 – initial set with manual markup (so-called seed-set), usually 10% of the full markup; B – number of images that can be signed in one cycle; T – total number of active iterations.

Thus, this condition ensures that the total amount of manual markup corresponds to the established budget. Equations (15) and (16) together form an optimisation problem with constraints: it is necessary to maximise the accuracy gain on rare classes without exceeding the available human resource. The proposed example selection strategy focused on rare categories (tail-aware sampling) demonstrates an advantage over random or purely entropy methods.

Comparison of Tail-Aware Active Self-Training strategy with other basic approaches

As part of the experimental study, the proposed Tail-Aware Active Self-Training strategy was compared with two basic approaches that reflect the lower and intermediate limits of the effectiveness of active training. The first basic scenario is Random, in which a fixed number of examples $B = 256$ are randomly selected at each iteration from a pool of unsigned images U. This approach does not consider either the level of uncertainty of the model or the frequency characteristics of classes, and therefore acts as a minimal control that allows assessing whether there is any benefit from using active learning.

The second basic option is the Uncertainty-only strategy, which is based on the classical entropy sorting approach proposed by B. Settles (2009). In this case, the images are ranked by the total entropy of the model's predictions, and the examples that the model is most uncertain about are added to the sample. However, this strategy ignores information about the imbalance in the class representation, which is especially important for objects that belong to rare categories. Ultimately, TAAST combines the entropy approach with the weight gain of rare classes (via a multiplier $\omega_{c(x)}$) and a semantic diversity mechanism based on clustering of normalised clip feature vectors. This approach helped to avoid excessive duplication of such personnel and ensured the maximum increase in information per unit of human resource. A consistent comparison of Random → Uncertainty-only → TAAST strategies illustrated the contribution of both the fact of active learning itself and the additional effect of taking into account the structure of the long tail (tail-aware logic).

Performance evaluation was carried out using the basic AP_rare metric, i.e., average accuracy only for objects with less than 10 examples in the initial training sample. The calculation was performed by the official LVIS/COCO API at 10 IOU (Intersection over Union) thresholds in the range of 0.50-0.95 in 0.05 increments, which ensured compatibility with previous studies in the field of long-tail detection. To ensure that the improvement in AP_rare is not accompanied by a degradation in overall accuracy, the mAP_overall metric was additionally recorded – the average accuracy for all classes using the same protocol. Label Efficiency (LE) was also evaluated – the percentage of manual markup saved compared to the full training set (for example, LE = 42% means using only 58% of real labels to achieve a given quality).

A complete three-cycle experiment was performed for each strategy under study ($T = 3$), where three fixed initial seed values were used: 21, 42, and 63. At the zero cycle stage ($T = 0$) the basic values of AP_rare and mAP were recorded, and then after each active cycle ($T = 1, 2, 3$) was evaluated on a validation subset. This step-by-step assessment allowed tracking the dynamics of learning and identifying at what stage the performance plateau is reached. After the third iteration was completed, the test part was opened once for the final measurement – this allowed adjusting to the test data. Average values and 95% confidence intervals were calculated using the equation:

$$\bar{x} \pm 1.96\,\sigma/\sqrt{3}, \quad (17)$$

where $\bar{x}$ – mean metric value; σ – standard deviation of three runs with different seeds.

As part of the evaluation of the effectiveness of active learning strategies, a comparative analysis of manual labour costs and computational time was carried out with a fixed budget for three active cycles of 256 frames each (a total of 768 frames on top of the initial seed set, which was 10% of the train part of the LVIS v1.0 dataset ≈ 10,000 images). Table 1 shows that the proposed TAAST strategy achieved the highest label efficiency (LE = 42%), reducing the need for manual annotation. Specifically, it retained 9% of human effort compared to Random and Unknown-only, which showed only 33% LE. In addition, TAAST demonstrated an advantage over qualitative metrics (AP_rare), proving the effectiveness of including weighting factors for rare classes and combining pseudo-markings with semantic clustering.

Table 1. Cost analysis

Strategy	Manual frames per cycle	Manual frames per 3 cycles	Label-efficiency
Random	256	768	33%
Uncertainty-only	256	768	33%
TAAST	256	768	42%

Source: compiled by the author based on the results of the experiment

As can be seen from the table, the advantage of TAAST is a combination of classified entropy, semantic clustering, and pseudo-markup, which allows not only to reduce human effort, but also to achieve higher AP_rare accuracy values in rare classes. Thus, the experimental results confirmed that TAAST provides a more economical use of the annotation budget without compromising the quality of the model, which is a key factor for deploying systems in real-world conditions with limited resources. Table 2 summarises the totals for all active learning strategies tested on LVIS v1.0 and nuImages-Imbalanced datasets. Accuracy in rare classes (AP_rare), overall quality (mAP), human markup performance (LE), and gain after three active learning cycles were evaluated.

Table 2. Analysis of various active learning strategies

Dataset / Method	AP_rare, start	AP_rare, final ± 95% CI	Δ AP_rare	mAP_overall, final	Label efficiency
LVIS Random	12.6	14.6 ± 0.32	+2.0	34.1	33%
LVIS Uncertainty	12.6	17.6 ± 0.25	+5.0	35.4	33%
LVIS TAAST	12.6	18.9 ± 0.20	+6.3	36.0	42%
nuImages Random	21.3	23.2 ± 0.34	+1.9	41.6	34%
nuImages Uncertainty	21.3	26.3 ± 0.29	+5.0	42.8	34%
nuImages TAAST	21.3	27.7 ± 0.25	+6.4	43.2	43%

Source: compiled by the author based on the results of the experiment

The results show a clear advantage of the TAAST strategy in all aspects considered. The increase in the AP_rare metric on both datasets was more than + 6 percentage points, exceeding the "net" uncertainty by about + 1.3 percentage points. This confirms the effectiveness of combining entropy selection with logarithmically weighted selection of rare classes. The overall mAP has also grown, which means that there is no degradation in common classes. In addition, Label Efficiency exceeded 42-43%, which indicates a significant reduction in the need for manual markup. Since the 95% confidence intervals between TAAST and Uncertainty-only do not overlap, the difference is statistically significant.

The results obtained confirmed the effectiveness of the TAAST strategy in the context of long-tail active self-learning tasks. Compared to classic uncertainty-based selection scenarios (Settles, 2009), TAAST provides a significantly higher increase in accuracy in rare categories (AP_rare), while maintaining or even improving the overall mAP. This indicates an effective integration of pseudo-labels and frequency weighting and an analysed reduction in the need for manual marking. The current results are consistent with the findings by K. Sohn *et al*. (2020), which showed that high-quality pseudomarking combined with self-learning can be effective, although they did not consider the class imbalance. TAAST extends this idea by adding adaptive weighting over the frequency of classification categories.

In addition, the results are consistent with a number of new approaches in long-tail detection and active learning. In particular, the Plug-and-Play Active Learning (PPAL) method (Yang *et al*., 2024) implements a two-step sampling scheme focused on sample diversity, which is easily integrated into standard detection pipelines without significant architectural changes and provides stable AP growth with minimal overhead. In the field of 3D detection, the Rare Example Mining (REM) approach, proposed by C.M. Jiang *et al*. (2022), addresses the intra-class long tail by purposefully selecting rare examples: the combination of data-centric and model-centric steps allows achieving performance close to fully marked models, with significantly less manual labels. The long-tail problem in unmanned driving is systematically formalised by the LT3D method presented by N. Peri *et al*. (2023): hierarchical loss and multimodal RGB + LiDAR Fusion have been shown to significantly improve the accuracy of rare classes (such as "stroller") by better distinguishing small objects. Ultimately, in the broader context of the open long tail of OLTR++, Z. Liu *et al*. (2022) proposed an integrated framework with dynamic meta-embedding and modular active learning that simultaneously covers the imbalance, few-shot, and open-set aspects — the results were confirmed on large ImageNet, Places, and MS1M sets.

Thus, the experimental results demonstrate that the TAAST strategy can combine the benefits of active and self-learning approaches in a single cycle. It allows significantly reducing the number of frames that require manual annotation, while not losing the overall accuracy of the model. It is important to note that the integration of frequency weighting directly affects the balance of the distribution of selected examples, which is crucial for improving rare classes. This suggests that not only architectural solutions, but also the process of forming a training set itself can become a key factor in improving the efficiency of object detection systems. Overall, the study not only confirmed the effectiveness of active learning as a concept, but also showed that the consideration of class frequency and semantic diversity allows for a much better balance between model quality and annotation costs. Thus, the proposed approach solves one of the key problems of active learning – the preference for frequent classes in the selection process – and offers a practical solution for problems with an imbalanced distribution that often occur in real-world conditions.

Conclusions

The proposed Tail-Aware Active Self-Training method confirmed the effectiveness of targeted and informative sampling in active self-learning tasks. Unlike classical strategies that focus only on entropy or randomness of choice, TAAST combines a rarity weighting factor in combination with an entropy estimation of model uncertainty, which allows prioritising frames with underrepresented classes. This approach provided an increase in average accuracy for rare objects by 6.3-6.4 percentage points, exceeding the results of both random and entropy active learning strategies. Using the 0.8 threshold for pseudo-marking helped to automatically include up to 50% of objects in the training set without the need for manual marking, which resulted in human resource savings of up to 43%. However, the quality of the model did not deteriorate, but on the contrary – it increased both at the level of rare classes and at the level of the overall average indicator. A key role in achieving high efficiency was played by the use of the CLIP model, which allows evaluating the semantic similarity of images without additional training, and clustering by the *k*-means method, which provided grouping scenes in seconds. This allowed avoiding duplication of frames and guarantee maximum diversity in the sample. It was shown that the model reaches a plateau of quality growth after the second iteration of active learning, which allows limiting further additional learning to 30-epoch cycles without losing productivity.

Thus, the method significantly speeds up the detector's self-learning, reduces computational costs, and minimises dependence on expensive manual annotations, making it suitable for use in real-world production environments with a limited budget. Prospects for further research are to extend the method to multimodal data sets, where text or sensory information is available in addition to images. It is also advisable to explore the possibility of adapting TAAST to video streams, where the time context can improve the quality of frame selection. In addition, an important area is the development of an automated mechanism for dynamically adjusting weight coefficients depending on changes in the statistics of marked-up data during the active cycle.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Ali, M.L., & Zhang, Z. (2024). The YOLO framework: A comprehensive review of evolution, applications, and benchmarks in object detection. *Computers*, 13(12), article number 336. doi: 10.3390/computers13120336.

[2] Bottou, L. (2012). Stochastic gradient descent tricks. In G. Montavon, G.B. Orr & K.R. Müller (Eds.), *Neural networks: Tricks of the trade*. *Lecture notes in computer science* (Vol. 7700, pp 421-436). Berlin: Springer. doi: 10.1007/978-3-642-35289-8_25.

[3] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 11621-11631). Seattle: IEEE/CVF. doi: 10.1109/CVPR42600.2020.01164.

[4] De Alvis, C., & Seneviratne, S. (2024). A survey of deep long-tail classification advancements. *ArXiv*. doi: 10.48550/arXiv.2404.15593.

[5] Duan, C.-L., Li, Y., Wei, X.-S., & Zhao, L. (2024). Longtail object detection pre-training: Dynamic rebalancing contrastive learning with dual reconstruction. In *38th conference on neural information processing systems (NeurIPS 2024)*. Vancouver: NeurIPS.

[6] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of Machine Learning Research*, 48, 1050-1059.

[7] Jocher, G., Chaurasia, A., & Qiu, J. (2023). *YOLOv5 and YOLOv8: A detailed comparison*. Retrieved from https://docs.ultralytics.com/models/yolov8/.

[8] Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535-547. doi: 10.1109/TBDATA.2019.2921572.

[9] Li, B., Yao, Y., Tan, J., Zhang, G., Yu, F., Lu, J., & Luo, Y. (2022). Improving long-tailed object detection with image-level supervision by multi-task collaborative learning. *ArXiv*. doi: 10.48550/arXiv.2210.05568.

[10] Li, Y., Wang, T., Kang, B., Tang, S., Wang, Ch., Li, J., & Feng, J. (2020). Overcoming classifier imbalance for long-tail object detection via balanced group softmax. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10991-11000). Seattle: IEEE. doi: 10.1109/CVPR42600.2020.01100.

[11] Qi, T., Xie, H., Li, P., Ge, J., & Zhang, Y. (2023). Balanced classification: A unified framework for long-tailed object detection. *IEEE Transactions on Multimedia*, 26, 3088-3101. doi: 10.1109/TMM.2023.3306968.

[12] Radford, A., *et al*. (2021). Learning transferable visual models from natural language supervision. In *38th international conference on machine learning (ICML 2021)* (pp. 8748-8763). Online Conference.

[13] Sener, O., & Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. In *ICLR 2018 conference track: 6th international conference on learning representation*. Vancouver: Vancouver Convention Center.

[14] Settles, B. (2009). *Active learning literature survey*. Madison: University of Wisconsin-Madison.
[15] Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., & Raffel, C. (2020). FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *NIPS'20: Proceedings of the 34th international conference on neural information processing systems* (pp. 596-608). Vancouver: NIPS.
[16] Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: Fully convolutional one-stage object detection. In *IEEE/CVF international conference on computer vision (ICCV)* (pp. 9626-9635). Seoul: IEEE. doi: 10.1109/ICCV.2019.00972.
[17] Wu, J., Chen, J., & Huang, D. (2022). Entropy-based active learning for object detection with progressive diversity constraint. In *IEEE/CVF conference on computer vision and pattern recognition* (CVPR) (pp. 9387-9396). New Orleans: IEEE. doi: 10.1109/CVPR52688.2022.00918.
[18] Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., & Liu, Z. (2021). End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 3060-3069). Montreal: IEEE. doi: 10.1109/ICCV48922.2021.00305.
[19] Yang, C., Huang, L., & Crowley, E.J. (2024). Plug-and-play active learning for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2024)* (pp. 17784-17793). Seattle: IEEE. doi: 10.1109/CVPR52733.2024.01684.
[20] Jiang, C.M., Najibi, M., Qi, C.R., Zhou, Y., & Anguelov, D. (2022). Improving the intra-class long-tail in 3D detection via rare example mining. In *Computer vision – ECCV 2022. Lecture notes in computer science* (Vol. 13670, pp. 155-172). Cham: Springer. doi: 10.1007/978-3-031-20080-9_10.
[21] Peri, N., Dave, A., Ramanan, D., & Kong, S. (2023). Towards long-tailed 3d detection. *Proceedings of Machine Learning Research,* 205, 1904-1915.
[22] Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., & Yu, S.X. (2022). Open long-tailed recognition in a dynamic world. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 46(3), 1836-1851. doi: 10.1109/TPAMI.2022.3200091.

Активне самонавчання для детекції об'єктів в умовах дисбалансованих даних: підхід TAAST

Дмитро Іванов
Аспірант
Державний університет «Житомирська політехніка»
10005, вул. Чуднівська, 103, м. Житомир, Україна
https://orcid.org/0000-0002-7386-4497

Анотація. У контексті дедалі ширшого розвитку та застосування комп'ютерного зору зростає потреба у зменшенні витрат на ручну розмітку даних, особливо в задачах виявлення рідкісних об'єктів за умов довгохвостого розподілу класів. Метою дослідження було підвищення ефективності визначення рідкісних категорій зображень через вдосконалення стратегії активного самонавчання. У роботі застосовано підхід Tail-Aware Active Self-Training, що базується на стратегічному відборі кадрів з урахуванням ентропії невпевненості, рідкісності класу та семантичного різноманіття в просторі ознак моделі Contrastive Language-Image Pretraining, з подальшим використанням псевдорозмітки за допомогою детектора You Only Look Once, версія 8. У результаті експериментів на наборах даних Large Vocabulary Instance Segmentation, версія 1.0 та nuImages-imbalanced запропонована стратегія забезпечила приріст точності AP_rare на 6,3–6,4 відсоткових пунктів у порівнянні з базовими підходами Random та Uncertainty Sampling. Загальна точність моделі при цьому не знизилась, а зросла до 36,0–43,2 % mAP залежно від датасету. Показник ефективності розмітки досягнув 42–43 %, що на 9-10 пунктів вище за конкурентні стратегії. Результати експерименту є статистично достовірними, оскільки інтервали довіри для метрики точності AP_rare у разі застосування методу Tail-Aware Active Self-Training не перетинаються з інтервалами для базових стратегій Random і Uncertainty-only. Це свідчить про те, що перевага даного методу не є випадковою, а підтверджена з високою ймовірністю. Отже, отримані результати продемонстрували надійність і стабільність запропонованого підходу: вже після двох активних ітерацій модель досягла плато продуктивності, що дозволило суттєво зменшити обчислювальні витрати. Практична цінність роботи полягає у створенні ефективного інструменту для автоматизованого розгортання моделей комп'ютерного зору в умовах обмеженого бюджету на розмітку

Ключові слова: машинне навчання; семантична кластеризація; псевдоанотація; вибірка за ентропією; балансування класів; комп'ютерний зір; оптимізація розмітки

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 05.08.2025 Revised: 03.11.2025 Accepted: 23.12.2025

UDC 004.4:004.42 DOI: 10.31649/vitce/3.2025.65

Methodology for designing memory-safe high-performance applications using layered resource isolation

Olha Krasnozhon*
Master
Academician Stepan Demianchuk International University of Economics and Humanities
33000, 4S Demianchuk Str., Rivne, Ukraine
https://orcid.org/0009-0008-0202-9575

Abstract. This study presented a design strategy – Layered Resource Isolation – that reconciled memory safety with high performance by enforcing three explicit tiers of lifetimes and checks: an ephemeral tier for short-lived temporaries, a verifiable tier guarded by structural and aliasing validation at transfer points, and a persistent tier with audited release. The objective was to elevate lifetime boundaries to first-class design elements while avoiding vendor-specific frameworks. Neutral exemplars preserved identical algorithms across baseline and layered variants: a parser and compiler front-end that transforms token streams into abstract syntax trees, a multi-level cache with coherent read-through behaviour, and blocked numerical kernels. The evaluation instrumented allocations, promotions, audited releases, and phase timings, and used paired runs across thirty independent seeds to compare safety incidents per ten million operations, median runtime, ninety-fifth and ninety-ninth percentile latencies, throughput, and peak resident memory. Results showed elimination of leaks, double frees, use-after-free, and invalid frees within the detection horizon in all layered variants, with a one-sided confidence bound placing the incident rate below 0.11 per ten million operations. Tail behaviour improved markedly: ninety-fifth percentiles decreased by 21.8-24.9% and ninety-ninth percentiles by 22.8-27.6% across exemplars and load regimes, peak resident memory fell by 10-16%, steady-state throughput rose by 0.6-4.1%, and median runtime overhead remained near 1-2%. Practically, the approach reduced allocator contention, enabled whole-program reasoning about ownership and aliasing, and converted rare, expensive recovery into predictable boundary validation, offering a replicable methodology for advanced systems software

Keywords: audited release; ownership and alias control; validation checkpoints; allocator and handle contracts; alias guards; typestate encodings

Introduction

The tension between memory safety and high performance remains a central constraint in systems software. Manual allocation and deallocation provide precise control yet introduce error pathways that manifest as leaks, dangling references, double frees, and subtle lifetime violations. Tracing garbage collection reduces explicit ownership burden but can impose non-deterministic pauses, increase cache churn, and obscure the moment when resources are reclaimed. Deterministic destructor-based patterns (RAII – Resource Acquisition Is Initialisation) grant predictable clean-up but localise reasoning to scope boundaries that may not reflect aliasing realities across modules. Under realistic workloads, these mechanisms can degrade throughput, inflate tail latency, and compromise correctness. A design-level methodology that elevates memory safety to a first-class concern while preserving speed is therefore warranted.

Layered Resource Isolation (LRI) is proposed as such a methodology that is suitable for open academic use. The central premise holds that many safeties failures stem from blurred lifetime boundaries and ad hoc promotion of data from transient scopes to long-lived structures. To counter this, LRI partitions memory and related resources into three tiers with clear contracts and checkpoints: an ephemeral tier for short-lived temporaries and scratch buffers with strictly bounded scope; a verifiable tier for state that must pass structural, aliasing, and invariance

Suggested Citation:
Krasnozhon, O. (2025). Methodology for designing memory-safe high-performance applications using layered resource isolation. *Information Technologies and Computer Engineering*, 22(3), 65-76. doi: 10.31649/vitce/3.2025.65

*Corresponding author

checks before any promotion; and a persistent tier for long-lived objects with stable identity, explicit transfer semantics, and audited release. R.N. Watson *et al.* (2025) argued that software ecosystems need standardise principles and measurements for memory safety to curb entire classes of defects. The position paper outlined baseline practices and stresses cross-project coordination so improvements persist beyond individual languages.

This stratification complements – while remaining distinct from – prevailing practices. Manual management emphasises control but lacks uniform validation boundaries; tracing collectors amortises reclamation but relaxes explicit ownership; RAII enforces deterministic clean-up yet depends on local scope reasoning that can mask cross-component aliasing. Compatibility with C/C++ libraries and common toolchains is maintained to enable incremental adoption without vendor lock-in or domain-specific frameworks. A. Fromherz & J. Protzenko (2024) formalised a pathway for compiling C into safe Rust with machine-checked semantics. The work showed how verified transformations can retain performance while inheriting Rust's ownership and lifetime guarantees. V. Astrauskas *et al.* (2022) presented Prusti, a verification toolchain for Rust that enables specification and automated checking of program properties. The project strengthened Rust's safety model by proving invariants that exceed what the borrow checker enforces.

The methodological contribution is framed as strategy and methods rather than invention. First, a vocabulary for lifetime governance is established – entry and exit contracts for each tier, typestate or equivalent static encodings where available, and lightweight runtime guards where static proof is impractical. Second, neutral exemplars are used to illustrate application without domain coupling: parser and compiler front-ends where tokens become abstract syntax trees through verifiable promotion; cache layering where lines migrate only via validated coherence steps; and numerical kernels where working sets remain in the ephemeral tier and results are promoted under explicit invariants. S. Amar *et al.* (2023) introduced Capability Hardware Extension to RISC-V for Internet of Things (CHERIoT), bringing capability-based spatial and temporal memory safety to embedded devices. The evaluation demonstrates fine-grained compartmentalisation with minimal footprint, suggesting feasibility in constrained environments. A.E. Michael *et al.* (2023) proposed MSWasm (Memory-Safe WebAssembly), enforcing memory-safe execution of unsafe code via WebAssembly isolation. The runtime establishes sound boundaries for legacy components without demanding extensive rewrites.

S. Xu *et al.* (2024) developed Condo to harden container isolation by protecting kernel permission metadata. The approach reduces escalation vectors by safeguarding critical authorisation paths. Several gaps motivate this inquiry. Existing literature and practice provided powerful individual mechanisms – ownership types, region disciplines, hazard pointers, epoch reclamation, and advanced allocators – yet guidance on composing these elements into a language-agnostic, design-time methodology remains limited. Moreover, prior work often isolates formal safety guarantees from performance characterisation, whereas engineering practice demands co-optimisation. D. Greenspan *et al.* (2024) presented LOaPP (Low-Overhead Protection for Persistent memory), a low-overhead scheme that protects persistent memory objects at rest. Results indicated that practical security policies can coexist with high performance in PM-backed systems. H. Huang *et al.* (2024) introduced vKernel, which gives each container private code and data spaces to tighten intra-kernel separation. Experiments showed improved isolation and reduced cross-container interference with modest overhead.

The objective of the present study was to articulate LRI as a transparent methodology – including design rules, validation checkpoints, and cross-tier contracts – and to examine its implications for safety and performance using neutral software exemplars. The research problem is articulated as follows: how can a layered lifetime discipline reduce memory misuse while preserving high throughput and predictable latency in production-grade codebases? The working hypothesis states that explicit tier boundaries with mandatory validation checkpoints reduce both incidence and impact of misuse – and that the resulting locality and deterministic reclamation improve tail behaviour despite modest checking overheads. A secondary hypothesis anticipates that whole-programme reasoning is strengthened when cross-tier transfers are promoted to first-class operations governed by contracts.

Materials and Methods

Methodology: Layered resource isolation

The study operationalised a design methodology – Layered Resource Isolation – to reconcile memory safety with high performance. LRI partitioned memory and related handles into three tiers with bounded lifetimes and mandatory checkpoints. The ephemeral tier hosted temporaries and scratch buffers whose arenas were created and torn down at natural phase boundaries. The verifiable tier retained the intermediate state subject to structural, bounds, and aliasing checks; only objects that satisfied these contracts were eligible for promotion. Before promotion, the object in the verifiable tier undergoes a gate check: its structural integrity and bounds are validated, the absence of ephemeral and overlapping aliases is confirmed, typestate and single ownership are verified with no writers in flight and no epoch skew, then the object is sealed, assigned a stable identity, recorded in the ownership table, and logged for audited release. The persistent tier encapsulated long-lived objects with stable identity, explicit transfer semantics, and audited release. Cross-tier movement was governed by compile-time typestate encodings where available and by lightweight runtime guards otherwise. Allocator and handle contracts defined entry/exit conditions for each tier and prohibited references that crossed lifetime boundaries unchecked. The method remained language- and vendor-neutral and avoided domain-specific workflows; it was framed

as a strategy and set of methods rather than an invention, ensuring free academic use.

Neutral exemplars and implementation materials

To isolate lifetime discipline from algorithmic effects, three neutral software exemplars were implemented twice each: a baseline using idiomatic manual allocation and RAII/region patterns, and an LRI variant that preserved identical algorithms and dataflow while routing all allocations and transfers through tier contracts. The exemplars were (i) a parser and compiler front-end that transformed token streams into abstract syntax trees; (ii) a cache layering stack with coherent read-through behaviour across two and three levels; and (iii) numerical kernels drawn from blocked dense linear algebra and transform routines. These three exemplars were selected as representative because they collectively span orthogonal memory access and aliasing patterns (pointer-rich AST (Abstract Syntax Tree) graphs, coherent multi-level cache lines, and regular blocked arrays), remain domain-neutral and reproducible without vendor dependencies, provide clear checkpoints for promotion contracts, and, by keeping algorithms identical between the baseline and the LRI variant while changing only the tier wiring, allow observed effects to be attributed specifically to the layered architecture. All implementations targeted a portable POSIX (Portable Operating System Interface) environment and a standard C/C++ toolchain, with no vendor frameworks or device-specific APIs. Instrumentation in both variants recorded allocations, deallocations, and timing at phase boundaries; the LRI variant additionally logged promotion attempts, alias-guard triggers, and audited releases. Safety was examined by running sanitiser-assisted builds of the baseline to surface leaks, double frees, use-after-free, and invalid frees, and by examining LRI contract outcomes (e.g., rejected promotions) that would have manifested as defects without tiering. Safety was assessed using Clang/LLVM AddressSanitizer (ASan) with leak detection (LSan) enabled, supplemented by UndefinedBehaviorSanitizer (UBSan; bounds, null, vptr, and pointer-overflow checks); ThreadSanitizer (TSan) was applied to the cache exemplar to exclude race confounders, whereas MemorySanitizer (MSan) was not employed due to deliberate scratch-buffer initialisation patterns.

Measurement protocol and statistical analysis

Workloads were designed to stress allocation and lifetime behaviour, independent of domain semantics. Arrival processes covered near-normal traffic, heavy-tailed allocation sizes, and bursty phases that induced allocator contention. Each configuration (subsystem×regime×variant) was executed across 30 independent seeds to enable paired comparisons and variance estimation. Seed variation was controlled by a deterministic Pseudo-Random Number Generator (PRNG) at the harness level: for each run, one of 30 unique seeds parameterised all stochastic components of the workloads, including inter-arrival timings under the latency regimes, allocation sizes and lifetimes, burst positions and durations, parser input permutations, cache access orders and tiebreakers, and initial tiling choices in numerical kernels; build and environment settings were held constant to ensure independence and reproducibility. Metrics included median runtime (p50), tail latencies (p95, p99), steady-state throughput, peak resident memory, and incident rates per operation; in addition, dispersion statistics (interquartile range – IQR) and the p99/p50 ratio were computed as auxiliary indicators of latency spread and tail heaviness.

Throughput was computed as subsystem-specific operations per second over a steady-state window (warm-up and cool-down excluded) in a portable POSIX C/C++ harness: for the parser, tokens successfully parsed into AST nodes per second; for the cache stack, completed get and put operations per second with both hits and misses counted; and for the numerical kernels, completed blocked-kernel tile updates per second. Per-seed values were then summarised as LRI-versus-baseline relative deltas to normalise for hardware. Code size was also compared as an overhead proxy: "lines of core code" were counted per exemplar and per variant (baseline vs LRI), including only hand-written kernel sources and excluding tests, harness utilities, build configuration, third-party libraries, generated files, blank lines, and comments, and the metric was reported both as absolute counts and as LRI-to-baseline deltas. Relative deltas were computed as LRI versus baseline for each seed to control for run-to-run variability. Statistical analysis followed a paired, non-parametric plan suitable for skewed latency distributions: the Wilcoxon signed-rank test was applied to per-seed deltas, with Holm-Bonferroni adjustment for families of comparisons across regimes and subsystems ($\alpha = 0.05$). For throughput and peak memory, 10,000-replicate bias-corrected bootstrap confidence intervals were computed on paired deltas.

For safety with zero observed incidents in LRI, one-sided Clopper-Pearson bounds estimated the upper rate consistent with the observations. Deterministic seeding and fixed build configurations ensured reproducibility. This protocol aligned with the Results section by contrasting structurally identical baselines against LRI-governed variants, attributing any safety improvements and tail-latency reductions to explicit lifetime layering, validation checkpoints, and audited release rather than to algorithmic changes. An ablation study was also conducted by disabling individual mechanisms (typestate checks, alias guards, ephemeral pooling, and audited release) while holding all other factors constant. Additionally, sensitivity tests were conducted by varying checkpoint frequency, ephemeral-pool size, and the batching policy for audited release.

Results

Benchmark setup, datasets, and measurement protocols

The evaluation tested whether the proposed LRI methodology – built around ephemeral, verifiable, and persistent tiers with explicit entry/exit contracts and mandatory validation checkpoints – reduced memory safety defects while

maintaining or improving performance. To ensure that any differences arose from lifetime governance rather than algorithmic changes, each exemplar subsystem was implemented twice with identical algorithms and dataflow: a baseline version using idiomatic manual allocation and RAII/region discipline, and an LRI version that preserved the same logic but routed all allocations, handles, and promotions through tier contracts.

Three neutral, domain-agnostic exemplars were chosen because they stress different forms of resource management without invoking vendor workflows or device-specific contexts. First, a parser and compiler front-end transformed token streams into abstract syntax trees (ASTs). In the baseline, tokens, stacks, and intermediate nodes coexisted in a single allocation regime; in the LRI variant, short-lived tokens and stacks resided in ephemeral pools; partially built AST nodes and candidate symbols were placed in the verifiable tier and were promoted only after structural checks (balanced subtrees, resolved ownership of lexemes); and finalised AST and symbol tables moved to the persistent tier with audited release. Second, a cache layering stack implemented a coherent, read-through policy across two and three levels. The baseline relied on conventional invalidation paths; in the LRI version, allocations for request contexts and hash probes were ephemeral, candidate lines and decoded metadata remained verifiable until coherence and alias checks passed, and only then were stable lines persisted, and demotion and deallocation proceeded through audited release. Third, numerical kernels (blocked dense linear algebra and transform routines) used the same loop ordering and tile sizes in both implementations; the LRI variant confined tile-local scratch to ephemeral pools, kept partial results verifiable until dimensionality and bounds invariants were asserted, and promoted only consolidated outputs and reusable plans.

A common harness collected counts of allocations, deallocations, promotions, demotions, and audited releases; sanitiser-assisted incident detection (for the baseline) covering leaks, double-frees, use-after-free, and invalid free; LRI contract outcomes (accepted vs. rejected promotions, alias-guard triggers, audited release results); and timing at subsystem phase boundaries with p50/p95/p99 latencies and throughput under steady and bursty regimes. Workloads covered near-normal interarrival, heavy-tailed allocation-size distributions, and burst phases introducing allocator contention. Each configuration was repeated across 30 independent seeds per regime to bound variance and support paired statistical comparisons. In Table 1, there were the exemplar subsystems and their baseline/LRI variants, core lines of code, total allocation counts, promotion and audited-release volumes, and the workload regimes exercised.

Table 1. Workloads, scale, and instrumentation coverage

Subsystem (variant)	Lines of core code	Allocations (×10⁶)	Promotions (×10⁶)	Audited releases (×10⁶)	Regimes exercised
Parser/Front-end (baseline)	7,420	182.6	–	–	Near-normal, heavy-tailed, bursty
Parser/Front-end (LRI)	8,105	176.9	94.2	61.1	Near-normal, heavy-tailed, bursty
Cache layering (baseline)	6,033	139.4	–	–	Near-normal, heavy-tailed, bursty
Cache layering (LRI)	6,612	131.7	72.8	48.0	Near-normal, heavy-tailed, bursty
Numerical kernels (baseline)	5,487	214.9	–	–	Near-normal, heavy-tailed, bursty
Numerical kernels (LRI)	5,998	205.5	118.6	79.3	Near-normal, heavy-tailed, bursty

Source: created by the author

The table establishes parity of algorithmic scope while revealing structural differences introduced by LRI. Code size grew modestly (8-11%) because contracts and manifest-level tier wiring were added; however, allocation counts fell consistently in LRI (-3.1% in the parser, -5.5% in the cache, -4.4% in numerics). This reduction reflects ephemeral pooling and fewer accidental long-lived clones after disciplined promotion. Promotions and audited releases – nonexistent in the baseline – quantify verifiable persistent flows and accountable deallocations: the parser executed 94.2 million promotions and 61.1 million audited releases, indicating that not all verifiable objects required eventual persistence and that persistent objects were reclaimed under audit rather than ad hoc free paths. The cache and numerics show similar shaping of resource lifecycles (72.8/48.0 and 118.6/79.3 million promotions/audited releases, respectively). Crucially, the regimes exercised are identical across variants, enabling paired comparisons in subsequent tables. The setup therefore met the study's requirement: same algorithms, same workloads, and additional lifetime structure.

Safety outcomes across tiers and transfers

The primary question was whether explicit tier boundaries with validation checkpoints would reduce or eliminate common memory-safety defects. Baseline variants were compiled and run under sanitiser assistance to surface latent issues that might not crash immediately; LRI variants logged contract outcomes and audited every persistent-tier release. In Table 2, there were aggregated sanitiser-detected

incident rates per ten million operations – leaks, double-frees, use-after-free, and invalid frees – together with LRI-specific counts of rejected promotions and audited-release failures across regimes.

Table 2. Safety incidents per 10 million operations (mean across seeds)

Subsystem/Regime	Leaks	Double-frees	Use-after-free	Invalid free	Rejected promotions	Audited release failures
Parser (baseline, near-normal)	1.6	0.2	0.9	0.4	–	–
Parser (baseline, heavy-tailed)	2.3	0.3	1.8	0.6	–	–
Parser (baseline, bursty)	2.9	0.4	2.1	0.7	–	–
Parser (LRI, all regimes)	0.0	0.0	0.0	0.0	3.7	0.0
Cache (baseline, near-normal)	0.9	0.1	0.5	0.2	–	–
Cache (baseline, heavy-tailed)	1.5	0.2	1.1	0.4	–	–
Cache (baseline, bursty)	1.8	0.3	1.3	0.5	–	–
Cache (LRI, all regimes)	0.0	0.0	0.0	0.0	2.4	0.0
Numerics (baseline, near-normal)	0.7	0.0	0.3	0.1	–	–
Numerics (baseline, heavy-tailed)	1.1	0.1	0.8	0.3	–	–
Numerics (baseline, bursty)	1.4	0.1	1.0	0.4	–	–
Numerics (LRI, all regimes)	0.0	0.0	0.0	0.0	2.9	0.0

Source: created by the author

The baseline exhibited non-zero rates of every sanitiser-trackable defect class, with magnitudes increasing under heavy-tailed and bursty regimes. For instance, parser leaks rose from 1.6 to 2.9 per 10 million operations between near-normal and bursty regimes (+ 81.3%), while use-after-free nearly doubled (0.9→2.1, + 133.3%). Cache and numerics showed the same pattern, albeit with smaller absolute values in numerics owing to more regular lifetimes. In stark contrast, LRI variants recorded zero leaks, double-frees, use-after-free, and invalid frees across all regimes within the detection horizon. Rejected promotions were treated as contract outcomes rather than defects; they indicate verifiable-tier objects that did not satisfy structural or aliasing contracts and therefore were not persisted. The observed rates – 3.7, 2.4, and 2.9 per 10 million operations in parser, cache, and numeric – amount to approximately 0.0037-0.0024% of attempted promotions and correspond to conditions that, in the baseline, correlate with the sanitiser incidents seen in the adjacent rows (e.g., promoting an AST node holding a borrowed pointer to an ephemeral token buffer, or persisting a cache line with incoherent metadata).

No audited release failures occurred, demonstrating that persistent-tier deallocation happened only when all owning handles had been revoked or transferred per contract. A one-sided 95% confidence bound on the LRI incident rate with zero observed events placed the upper bound below 0.11 per 10 million operations, at least an order of magnitude under baseline means for the same subsystems and regimes. The directionality is consistent: when lifetime governance is explicit, entire defect classes disappear rather than merely decline. Mechanistically, the elimination follows directly from the tier semantics. Ephemeral pools are torn down wholesale at natural block boundaries, so "missed free on an error path" cannot accumulate into leaks; verifiable-tier promotion gates prevent any persistent object from holding pointers into ephemeral memory; alias guards catch stale handles before invalidation; audited release ensures that finalisation order is checked against ownership and alias maps, turning silent hazards into contract failures at the boundary rather than undefined behaviour deep in the program.

Latency, throughput, and peak memory footprint

The secondary question was whether LRI's contracts and audits, together with tier-aware allocations, preserved or improved performance metrics – especially tail latency – without imposing unacceptable median overhead. Timing probes were placed at phase boundaries in each subsystem, and results were summarised as relative deltas of the LRI variant against its baseline counterpart; negative deltas indicate reductions relative to baseline. In Table 3, there were relative deltas of the LRI implementation versus the baseline for median runtime, p95/p99 latency, throughput, and peak memory under near-normal, heavy-tailed, and bursty loads.

Table 3. Election and commit time percentiles (ms)

Subsystem/Regime	Median runtime Δ (p50)	p95 latency Δ	p99 latency Δ	Throughput Δ	Peak memory Δ
Parser (near-normal)	+1.3%	-22.4%	-24.1%	+0.9%	-10.7%
Parser (heavy-tailed)	+1.6%	-24.9%	-27.6%	+2.8%	-12.9%
Parser (bursty)	+1.7%	-23.1%	-26.3%	+3.2%	-13.8%
Cache (near-normal)	+1.2%	-21.8%	-23.2%	+1.1%	-10.1%
Cache (heavy-tailed)	+1.4%	-23.7%	-26.8%	+3.7%	-14.5%
Cache (bursty)	+1.5%	-22.6%	-25.4%	+4.1%	-15.9%

Table 3. Continued

Subsystem/Regime	Median runtime Δ (p50)	p95 latency Δ	p99 latency Δ	Throughput Δ	Peak memory Δ
Numerics (near-normal)	+1.1%	-21.9%	-22.8%	+0.6%	-10.3%
Numerics (heavy-tailed)	+1.4%	-23.4%	-25.1%	+2.4%	-11.6%
Numerics (bursty)	+1.6%	-22.7%	-24.9%	+2.9%	-12.1%

Source: created by the author

The median runtime overhead remained around the design target of ~1.5% (+1.1 to +1.7%), while tail latencies improved substantially: p95 decreased by 21.8-24.9% and p99 by 22.8-27.6% across all subsystems and regimes. Throughput modestly increased, with the largest gains under heavy-tailed and bursty conditions (+2.4 to +4.1%), exactly where the baseline is most vulnerable to allocator contention and long-path recoveries. Peak memory declined by ~10-16%, a direct consequence of ephemeral pooling (scope-wide teardown), disciplined promotion (fewer unnecessary clones), and audited release (preventing lingering retention). The p99 improvement is particularly instructive. In the parser, late discovery of malformed intermediate structures in the baseline triggered expensive recovery and piecemeal teardown, inflating the tail; in the LRI variant, invalid intermediate states failed fast in the verifiable tier and were discarded cheaply, keeping the slow path out of the hot loop. In the cache stack, explicit alias guards removed serialised read-modify-write episodes caused by stale handles, which under bursty writeback amplified into p99 spikes; with guards, those episodes were prevented or resolved early, flattening the upper tail (Hardin, 2023). In numerical kernels, ephemeral pooling improved locality by keeping temporaries in compact arenas with predictable lifetimes, reducing allocator traffic; promotion gates stopped partial tiles from polluting persistent structures and triggering compensatory passes, again cutting the tail.

Dispersion statistics (not shown in the table but computed from the same runs) reinforce the picture: the interquartile range of operation latencies shrank by 9-13% across all subsystems, while the p99/p50 ratio – an informal tail-heaviness index – declined by 21-25%. The observed median overhead reflected work performed by contract checks and audits; by shifting error handling from sporadic recovery to early, low-cost validation, less time was spent in slow paths. Heavy-tailed regimes showcase this effect: the LRI variant improved p95/p99 and converted a portion of the baseline's throughput variability into stable gains, as seen in +3.7% (cache) and +2.8% (parser) throughput under heavy-tailed loads. Peak memory reductions provide a second-order performance benefit. Smaller footprints improve cache residency of hot data (e.g., near-term AST nodes, cache metadata, and tile descriptors), which feeds back into latency stability. The -14.5% peak reduction in the cache under heavy-tailed regimes aligns with the strongest throughput improvement in that row (+3.7%), suggesting that lifetime discipline acts both as a correctness guard and as a soft capacity optimiser. Statistical checks on paired runs across 30 seeds indicate that p99 improvements remained significant after Holm-Bonferroni adjustment ($\alpha = 0.05$), and bootstrap confidence intervals for throughput and peak-memory deltas excluded zero for heavy-tailed and bursty regimes. Near-normal regimes showed smaller throughput gains (0.6-1.1%) with confidence intervals brushing zero in some seeds, which is expected when slow-path avoidance matters less.

Ablations, sensitivity, and threats to validity

The findings support the working hypothesis: explicit tier boundaries with mandatory validation checkpoints eliminated sanitiser-detectable memory violations and simultaneously improved tail latency (p95 -21.8% to -24.9%, p99 -22.8% to -27.6%) at a small median overhead of ~1-2% with a throughput uptick; moreover, the low rate of rejected promotions together with zero faults in auditable reclamation is consistent with the secondary hypothesis that promoting cross-tier transfers to first-class, contract-governed operations strengthens whole-program reasoning. To isolate which parts of LRI carry the observed benefits, an ablation study disabled one mechanism at a time within otherwise identical LRI implementations: No-Typestate (contracts evaluated only at runtime), No-Alias-Guards (no cross-tier alias checks), No-Pooled-Ephemeral (ephemeral allocations drawn from the general allocator), and No-Audited-Release (persistent deallocation without final audits). Results are expressed relative to full LRI; positive values in latency indicate worse performance than full LRI, while "safety incidents" report newly observed defects per 10 million operations. In Table 4, there were ablation results showing, for each removed mechanism, the additional safety incidents and the change in p99 latency, peak memory, and throughput relative to full LRI.

Table 4. Ablations: effect on key outcomes (relative to full LRI)

Ablation	Safety incidents (per 10M ops)	p99 latency Δ vs. LRI	Peak memory Δ vs. LRI	Throughput Δ vs. LRI
No-Typestate	+0.04 (verifiable failures surfaced late)	+3.1%	+0.8%	-0.5%
No-Alias-Guards	+0.09 (occasional stale-handle misuse)	+4.7%	+0.6%	-0.9%
No-Pooled-Ephemeral	+0.00	+2.6%	+4.3%	-1.4%
No-Audited-Release	+0.00 (no immediate errors)	+0.8%	+2.1%	-0.3%

Source: created by the author

Removing compile-time encodings (No-Typestate) preserved correctness through runtime contracts but shifted some checks later, increasing p99 by 3.1% relative to full LRI and introducing a small but measurable incident rate (+0.04 per 10 million operations) in code paths with deep verifiable lifetimes. Disabling alias guards (No-Alias-Guards) had the most pronounced safety impact among ablations: stale-handle misuse reappeared at +0.09 per 10 million operations, and p99 degraded by 4.7%, underscoring that cross-tier alias hygiene is a key determinant of tail stability. Eliminating ephemeral pooling (No-Pooled-Ephemeral) did not add incidents in the measured horizon but increased peak memory by 4.3% and reduced throughput by 1.4%, tying the footprint advantage to pooled lifetimes.

Removing audited release (No-Audited-Release) produced no immediate sanitiser findings in the experiment window, but peak memory rose by 2.1% and throughput slipped slightly, indicating latent risk and soft capacity loss as unreclaimed objects accumulate for longer. These ablations collectively attribute benefits to distinct components: typestate-like encodings primarily dampen tails by pulling checks forward; alias guards directly prevent the most insidious class of misuse (stale references crossing invalidation); ephemeral pooling drives footprint and allocator contention; and audited release enforces long-horizon hygiene that may not manifest as acute failures in short runs but guards against drift in persistent stores.

Two tuning dimensions were explored. First, checkpoint frequency in the verifiable tier varied from "on promotion only" to "on promotion and again after every N operations" for $N \in \{103,104\}$. The additional checks improved margins under contrived heavy-tailed bursts, reducing the probability that long-lived verifiable objects accumulated inconsistent state; the trade-off was a minor erosion of tail improvements (p99 benefits shrank by 1-2 percentage points) and < 0.3% added median overhead. For the studied exemplars, "on promotion only" sufficed, with higher frequencies recommended when verifiable objects span many epochs by design.

Second, tier granularity controlled ephemeral pool sizes and batching policy for audited releases. Pools $\leqslant 64$ KiB lowered peak memory slightly but increased allocator traffic and fragmentation, reducing throughput by ~0.8%. Pools $\geqslant 1$ MiB boosted numeric-kernel throughput by ~0.4% via better temporal locality but produced transient peaks during bursts. A 256-512 KiB range balanced these effects across subsystems. Audited releases performed best when aligned with natural quiescence points (phase boundaries); overly coarse batching risked short-lived over-retention and measurable p95 inflation. Although the exemplars were chosen for breadth – graph construction and sharing (parser), coherence and alias hygiene (cache), and regular temporaries and locality (numerics) – the study cannot claim universal coverage. Different languages and toolchains will vary in how easily typestate-like encodings are expressed; nevertheless, the core contracts and audits translate to any environment that can enforce entry/exit checks and ownership transfer semantics.

Experiments were long enough to capture steady state and bursts but did not simulate months-long uptimes; in practice, the value of audited release is expected to compound over extended horizons. Taken together, these observations delineate the boundary conditions of the approach: once lifetime governance is explicit, allocator choice becomes a second-order factor, and the principal effects persist across workloads. Sensitivity knobs (checkpoint frequency, ephemeral-pool sizing, and release-batching) trade minor overheads for predictable stability, while ablations attribute most safety and tail improvements to alias guards and early validation. Accordingly, tier contracts and audited release are treated as the primary deployment levers, with allocator selection and granular tuning reserved for workload-specific shaping. The next section synthesises these implications, relating the measured effects to locality, contention, and fault containment, and distilling practical guidance for adoption.

Discussion

The study demonstrated that elevating lifetime boundaries and transfers to first-class design elements – ephemeral arenas for short-lived temporaries, a verifiable tier gated by structural and aliasing checks, and a persistent tier with audited release – suppressed sanitiser-detectable memory-safety defects to zero across all exemplars while improving tail behaviour and shrinking peak footprint. The median overhead remained near one to two per cent, yet p95/p99 latencies declined by roughly a quarter, and peak memory fell by about one-tenth to one-sixth. Ablations indicated that early checks encoded via typestate primarily dampened the upper tail, cross-tier alias guards prevented the most damaging stale-handle paths, pooled ephemeral allocation reduced allocator contention and fragmentation, and audited release contributed to long-horizon hygiene. Taken together, these results supported the central claim that explicit lifetime governance relocates error handling from rare, expensive, slow paths to cheap, predictable boundary validations.

The observed reductions in tail latency were consistent with evidence that memory-hierarchy discipline and working-set control dominate long-path behaviour under contention. M. Vaithianathan (2025) surveyed cache – DRAM (Dynamic Random Access Memory) hierarchies for high-performance workloads and emphasised that balanced placement and NUMA (Non-Uniform Memory Access)-aware policies sustain throughput at scale; by tightening promotions and pruning accidental persistence, the methodology reduced remote traffic amplification that typically harms p99 under load. N.D. Clerigo & J. Teleron (2025) compared allocator and reclamation strategies across systems, noting that lifetime misfit and fragmentation inflate variance; the arena-based ephemeral tier and refusal to persist unverifiable objects matched that prescription and explained the measured shrinkage of

dispersion. J.R.C. Jalaman & J.I. Teleron (2024) examined paging, caching, and allocation under realistic workloads and showed how locality decisions interact with OS policy; the footprint reductions recorded here plausibly compounded those effects, particularly in heavy-tailed regimes where allocator pressure and paging sensitivity align.

Runtime interference and resource sharing are well-known drivers of unstable tails, and the results under bursty regimes aligned with that literature. J. Kim & K. Lee (2020) analysed I/O resource isolation in serverless runtimes for data-intensive functions and identified interference channels that destabilise throughput; by limiting transient growth and preventing cross-tier aliasing, the methodology reduced time spent in allocator and cache slow paths and thereby reduced sensitivity to such interference. R. Bazuku *et al.* (2023) outlined operating-system trends toward security-aware kernels and lightweight isolation; lifetime discipline at the application layer complemented these shifts by shrinking the volume of undefined behaviour presented to the kernel. R. Abuleil *et al.* (2023) proposed a composite security blueprint for virtualised environments and reported reduced cross-tenant risk; the boundary contracts reported here further narrowed intra-process fault surfaces that virtualisation alone does not address. K. Sharma & P. Khurana (2025) synthesised container hardening patterns centred on least privilege and continuous verification; the audited-release rule and promotion gates embodied a similar verification stance within the process, providing a natural seam for external policy to target. M. Waseem *et al.* (2025) underscored portability and policy consistency in multi-cloud containerisation; because the tiering rules were vendor-neutral and encoded at design time, they fit that portability requirement. A.Y. Wong *et al.* (2023) mapped the container threat landscape and emphasised orchestrator policy and supply-chain assurance; while these concerns are orthogonal to memory discipline, eliminating intra-process misuse closed many avenues that such threats exploit.

At service boundaries, security guidance stressed identity, policy enforcement, and observability. M. Kothapalli (2021) delineated microservices practices along precisely those axes; within that framing, promotion contracts acted as a policy enforcement point that refused unsafe object sharing across interfaces, and audited release created observable, testable lifecycle events rather than ad hoc frees dispersed through the codebase. The empirical disappearance of double frees and use-after-free in the present study indicated that this internal policy materially simplified external enforcement by reducing undefined behaviour at the call boundary.

Fine-grained compartmentalisation inside a process provided another lens to interpret the ablations. D. Adak *et al.* (2025) introduced SpecMPK (Memory Protection Keys) and showed that speculative permission updates reduced the switching overhead of in-process isolation; the upper-tail erosion observed when alias guards were removed mirrored their conclusion that making boundary checks cheap and frequent is essential to practicality. M. Untergug-genberger *et al.* (2024) presented TME-Box (Total Memory Encryption), which attached per-domain encryption to partition data within an address space at modest cost; the methodology here achieved similar separation effects at the software-contract level, with comparable overhead magnitudes. K.D. Duy *et al.* (2023) proposed Capacity, a capability system that enforces fine-grained memory and API access; the refusal to promote unverifiable objects and the prohibition of cross-tier dangling references mirrored capability checks and suggested that software contracts and hardware capabilities are additive rather than exclusive. B. Joseph & R. Kavitha (2025) present radiation-hardened SRAM (Static random-access memory) with in-memory computing to tolerate soft errors in safety-critical contexts; while orthogonal to software lifetime governance, this hardware approach complements LRI by letting verifiable-tier checkpoints and audited promotions confine and discard corrupted states before they reach the persistent tier, thereby limiting blast radius without compromising throughput.

Language design also informed the interpretation of safety and tail stability. W. Bugden & A. Alahmar (2022) empirically compared languages and concluded that memory and concurrency choices materially shift defect rates and runtime costs; the present methodology replicated a portion of those benefits without a language migration by encoding lifetime rules and ownership transfers explicitly. T. Weis *et al.* (2019) introduced Fyr as a systems language pursuing memory and thread safety by construction; that direction was complementary, and the current findings provided a bridge for teams that could not adopt a new language wholesale. N. Yoshimura *et al.* (2024) proposed TECS/Rust (TOPPERS Embedded Component Systems/Rust) to enable componentised memory-safe composition with static checks in constrained environments; the verifiable-to-persistent promotion pattern mirrored such component contracts and suggested that the methodology could be encoded natively where stronger types are available.

A. Partap & D. Boneh (2022) designed an efficient memory-tagging scheme to detect spatial and temporal errors; the explicit boundaries reported here created natural cut-points where tag violations would surface early, simplifying diagnosis and limiting propagation. M. Gross *et al.* (2021) showed that malicious FPGA-SoC (Field-Programmable Gate Array – System on Chip) hardware could subvert isolation; although such attacks lie outside the software scope of lifetime governance, the absence of intra-process misuse reduced the exploitability of post-breach conditions and provided seams where attestation hooks can be anchored. System heterogeneity and energy-aware scheduling further contextualised the tail findings. J. Kim *et al.* (2024) proposed proactive boosting, saying that the anticipated workload needs to balance responsiveness and energy across accelerators; the more predictable phase behaviour produced by disciplined promotions and audited teardowns offered cleaner signals to such schedulers, making boosts

timelier and avoiding surprise stalls from allocator slow paths. M. Țălu (2025) compared WebAssembly runtimes along performance and sandbox strength and identified integration trade-offs; the contract-based design here fit naturally above a sandbox, clarifying lifetime even when executing in a managed module and easing the safe integration of native components.

From an integration standpoint, multi-layered system design emphasised standardised interfaces to reduce emergent complexity. R. Vadisetty (2024) discussed multi-layered technologies that enable interoperability across heterogeneous stacks and argued that standardised seams improve reliability; the tier boundaries and transfer semantics formalised in this methodology provided precisely such seams at the memory-semantics level, enabling independent teams to reason about ownership and lifetime without relying on shared tribal knowledge. I. Sañudo *et al.* (2020) reviewed memory constraints as central to mission-critical reliability and linked endurance, bandwidth, and security to predictable behaviour; while the evaluation here stayed in neutral software exemplars, the measured narrowing of latency dispersion and the elimination of memory incidents resonated with that requirement for predictability in critical contexts. Overall, the findings demonstrate that explicitly governed lifetime tiers not only eliminate memory-safety defects but also enhance predictability, efficiency, and integration readiness across heterogeneous systems, bridging the gap between low-level safety mechanisms and high-level architectural reliability.

Conclusions

This study has introduced Layered Resource Isolation as a design methodology that reconciles memory safety with high performance by governing object lifetimes across three explicit tiers – ephemeral, verifiable, and persistent – each with bounded scope, entry/exit contracts, and mandatory validation checkpoints. Using neutral exemplars (parser/compiler front-ends, cache layering, numerical kernels) implemented with identical algorithms in baseline and LRI variants, it has been established that the methodology eliminated sanitiser-detectable leaks, double frees, use-after-free, and invalid frees within the detection horizon. Quantitatively, long-tail latency improved by roughly one quarter (p95: -21.8% to -24.9%; p99: -22.8% to -27.6%), peak resident memory declined by ~10-16%, and steady-state throughput increased modestly (+0.6% to +4.1%) and remained at +1.1% to +1.7%.

These findings indicate that promoting lifetime boundaries and transfers to first-class design elements displaces rare, expensive recovery paths with cheap, predictable boundary checks. Typestate-like encodings brought validation earlier in the pipeline, cross-tier alias guards prevented stale-handle misuse, pooled ephemeral arenas reduced allocator contention, and audited release ensured long-horizon hygiene. In aggregate, LRI provided whole-program reasoning about ownership and aliasing without dependence on vendor stacks or domain-specific workflows and remained compatible with C/C++ libraries and common toolchains. Conservative defaults were applied in the evaluation: promotion-time verification, ephemeral pools sized 256 to 512 KiB, and release batching aligned to phase boundaries; all measurements were taken under these settings. Accordingly, within the studied scope, the results align with the working (and secondary) hypotheses: a disciplined separation of lifetimes with validation gates reduces the probability and impact of memory-management errors and, via locality and deterministic reclamation, stabilises tail latencies, while granting cross-tier transfers first-class, contract-based status reinforces whole-program reasoning.

Overall, LRI has been shown to elevate memory safety to a first-class design property while preserving, and often improving, performance, offering an immediately adoptable strategy for advanced systems software. Limitations included the finite breadth of exemplars, seed-bounded execution windows, and allocator diversity not exhaustively sampled. Future work should extend assessment to months-long runs, additional languages and compilers, specialised allocators, and richer concurrency patterns; integrate static verification pipelines to auto-synthesise contracts; and develop tooling that scaffolds tier wiring and audit instrumentation.

Acknowledgements

None.

Funding

The study was not funded.

Conflict of Interest

None.

References

[1] Abuleil, R., Murrar, S., & Shkoukani, M. (2023). An enhanced approach for realizing robust security and isolation in virtualized environments. *International Journal of Advanced Computer Science and Applications*, 14(11), article number 141129. doi: 10.14569/ijacsa.2023.0141129.

[2] Adak, D., Zhou, H., Rotenberg, E., & Awad, A. (2025). SpecMPK: Efficient in-process isolation with speculative and secure permission update instruction. In *2025 IEEE international symposium on high performance computer architecture (HPCA)* (pp. 394-408). Las Vegas: Institute of Electrical and Electronics Engineers. doi: 10.1109/HPCA61900.2025.00039.

[3] Amar, S., *et al*. (2023). CHERIoT: Complete memory safety for embedded devices. In *MICRO '23: Proceedings of the 56th annual IEEE/ACM international symposium on microarchitecture* (pp. 641-653). New York: Association for Computing Machinery. doi: 10.1145/3613424.3614266.

[4] Astrauskas, V., Bílý, A., Fiala, J., Grannan, Z., Matheja, C., Müller, P., Poli, F., & Summers, A.J. (2022). The prusti project: Formal verification for rust. In J.V. Deshmukh, K. Havelund & I. Perez (Eds.), *Proceeding of the 14th international symposium "NASA formal methods"* (pp. 88-108). Cham: Springer. doi: 10.1007/978-3-031-06773-0_5.
[5] Bazuku, R., Anab, A., Gyemerah, S., & Daabo, M.I. (2023). An overview of computer operating systems and emerging trends. *Asian Journal of Research in Computer Science*, 16(4), 161-177. doi: 10.9734/ajrcos/2023/v16i4380.
[6] Bugden, W., & Alahmar, A. (2022). The safety and performance of prominent programming languages. *International Journal of Software Engineering and Knowledge Engineering*, 32(5), 713-744. doi: 10.1142/s0218194022500231.
[7] Clerigo, N.D., & Teleron, J. (2025). A comparative study of memory management techniques and their optimization strategies. *International Journal of Advanced Research in Arts, Science, Engineering & Management*, 12(1), 39-50.
[8] Duy, K.D., Cho, K., Noh, T., & Lee, H. (2023). Capacity: Cryptographically-enforced in-process capabilities for modern ARM architectures. In *CCS'23: Proceedings of the 2023 ACM SIGSAC conference on computer and communications security* (pp. 874-888). New York: Association for Computing Machinery. doi: 10.1145/3576915.3623079.
[9] Fromherz, A., & Protzenko, J. (2024). Compiling C to safe rust, formalized. *ArXiv*. doi: 10.48550/arXiv.2412.15042.
[10] Greenspan, D., Mustafa, N.U., Delgado, A., & Bramham, C. (2024). LOaPP: Improving the performance of persistent memory objects via low-overhead at-rest PMO protection. In *2024 International symposium on secure and private execution environment design (SEED)* (pp. 131-142). Orlando: Institute of Electrical and Electronics Engineers. doi: 10.1109/SEED61283.2024.00023.
[11] Gross, M., Jacob, N., Zankl, A., & Sigl, G. (2021). Breaking TrustZone memory isolation and secure boot through malicious hardware on a modern FPGA-SoC. *Journal of Cryptographic Engineering*, 12(2), 181-196. doi: 10.1007/s13389-021-00273-8.
[12] Hardin, D. (2023). Hardware/software co-assurance for the RUST programming language applied to Zero Trust architecture development. *ACM SIGAda Ada Letters*, 42(2), 55-61. doi: 10.1145/3591335.3591340.
[13] Huang, H., Wang, H., Rao, J., Wu, S., Fan, H., Yu, C., Jin, H., Suo, K., & Pan, L. (2024). VKernel: Enhancing container isolation via private code and data. *IEEE Transactions on Computers*, 73(7), 1711-1723. doi: 10.1109/tc.2024.3383988.
[14] Jalaman, J.R.C., & Teleron, J.I. (2024). Optimizing operating system performance through advanced memory management techniques: A comprehensive study and implementation. *Engineering and Technology Journal*, 9(5), 4137-4143. doi: 10.47191/etj/v9i05.33.
[15] Joseph, B., & Kavitha, R. (2025). A radiation hardened in-memory computing SRAM for soft error tolerance in safety critical applications. *AEU – International Journal of Electronics and Communications*, 202, article number 156017. doi: 10.1016/j.aeue.2025.156017.
[16] Kim, J., & Lee, K. (2020). I/O resource isolation of public cloud serverless function runtimes for data-intensive applications. *Cluster Computing*, 23(3), 2249-2259. doi: 10.1007/s10586-020-03103-4.
[17] Kim, J., Lee, G., & Choi, H. (2024). Energy-efficient heterogeneous computing via normalized performance based proactive boost for embedded artificial intelligence. In *2024 IEEE international conference on consumer electronics (ICCE)* (pp. 1-6). Las Vegas: Institute of Electrical and Electronics Engineers. doi: 10.1109/ICCE59016.2024.10444347.
[18] Kothapalli, M. (2021). Securing microservices architecture: Best practices and challenges. *Journal of Scientific and Engineering Research*, 8(10), 187-192. doi: 10.5281/zenodo.12772079.
[19] Michael, A.E., *et al*. (2023). MSWasm: Soundly enforcing memory-safe execution of unsafe code. *Proceedings of the ACM on Programming Languages*, 7(POPL), 425-454. doi: 10.1145/3571208.
[20] Partap, A., & Boneh, D. (2022). Memory tagging: A memory efficient design. *ArXiv*. .doi: 10.48550/arXiv.2209.00307.
[21] Sañudo, I., Cortimiglia, P., Miccio, L., Solieri, M., Burgio, P., Di Biagio, C., Felici, F., Nuzzo, G., & Bertogna, M. (2020). The key role of memory in next-generation embedded systems for military applications. In P. Ciancarini, M. Mazzara, A. Messina, A. Sillitti & G. Succi (Eds.), *Proceedings of 6th international conference in software engineering for defence applications* (pp. 275-287). Cham: Springer. doi: 10.1007/978-3-030-14687-0_25.
[22] Sharma, K., & Khurana, P. (2025). A deep dive into container security challenges, strategies, and solutions. In *Proceedings of the international conference on recent advances in artificial intelligence for sustainable development (RAISD 2025)* (pp. 484-495). Dordrecht: Atlantis Press. doi: 10.2991/978-94-6463-787-8_38.
[23] Ţălu, M. (2025). A comparative study of WebAssembly runtimes: performance metrics, integration challenges, application domains, and security features. *Archives of Advanced Engineering Science*, 1-13. doi: 10.47852/bonviewaaes52024965.
[24] Unterguggenberger, M., Lamster, L., Schrammel, D., Schwarzl, M., & Mangard, S. (2024). TME-box: Scalable in-process isolation through intel TME-MK memory encryption. In *Network and distributed system security symposium 2025: NDSS 2025* (pp. 1-16). San Diego: Network and Distributed System Security. doi: 10.14722/ndss.2025.240277.

[25] Vadisetty, R. (2024). Multi layered cloud technologies to achieve interoperability in AI. In *2024 international conference on intelligent computing and emerging communication technologies (ICEC)* (pp. 1-5). Guntur: Institute of Electrical and Electronics Engineers. doi: 10.1109/ICEC59683.2024.10837471.

[26] Vaithianathan, M. (2025). Memory hierarchy optimization strategies for high-performance computing architectures. *International Journal of Emerging Trends & Technology in Computer Science*, 6(1), 24-35. doi: 10.63282/3050-9246.IJETCSIT-V6I1P103.

[27] Waseem, M., Ahmad, A., Liang, P., Akbar, M.A., Khan, A.A., Ahmad, I., Setälä, M., & Mikkonen, T. (2025). Containerization in multi-cloud environment: Roles, strategies, challenges, and solutions for effective implementation. *Journal of Systems and Software*, 230, article number 112558. doi: 10.1016/j.jss.2025.112558.

[28] Watson, R.N., *et al*. (2025). It is time to standardize principles and practices for software memory safety. *Communications of the ACM*, 68(2), 40-45. doi: 10.1145/3708553.

[29] Weis, T., Waltereit, M., & Uphoff, M. (2019). Fyr: A memory-safe and thread-safe systems programming language. In *SAC '19: Proceedings of the 34th ACM/SIGAPP symposium on applied computing* (pp. 1574-1577). New York: Association for Computing Machinery. doi: 10.1145/3297280.3299741.

[30] Wong, A.Y., Chekole, E.G., Ochoa, M., & Zhou, J. (2023). On the security of containers: Threat modeling, attack analysis, and mitigation strategies. *Computers & Security*, 128, article number 103140. doi: 10.1016/j.cose.2023.103140.

[31] Xu, S., Wang, Y., Lei, L., Sun, K., Jing, J., Ma, S., Wang, J., & Huang, H. (2024). Condo: Enhancing container isolation through kernel permission data protection. *IEEE Transactions on Information Forensics and Security*, 19, 6168-6183. doi: 10.1109/tifs.2024.3411915.

[32] Yoshimura, N., Oyama, H., & Azumi, T. (2024). TECS/Rust: Memory-safe component framework for embedded systems. In *2024 IEEE 27th international symposium on real-time distributed computing (ISORC)* (pp. 1-11). Tunis: Institute of Electrical and Electronics Engineers. doi: 10.1109/ISORC61049.2024.10551370.

Методологія проектування безпечних для пам'яті високопродуктивних застосунків з використанням багаторівневої ізоляції ресурсів

Ольга Красножон
Магістр
Міжнародний університет економіки та гуманітарних наук імені академіка Степана Дем'янчука
33000, вул. С. Дем'янчука, 4, м. Рівне, Україна
https://orcid.org/0009-0008-0202-9575

Анотація. У цьому дослідженні представлено стратегію проектування – багаторівневу ізоляцію ресурсів, яка поєднує безпеку пам'яті з високою продуктивністю шляхом застосування трьох явних рівнів терміну служби та перевірок: ефемерний рівень для короткочасних тимчасових ресурсів, рівень, що перевіряється, захищений структурною та аліасинговою перевіркою в точках передачі, та постійний рівень з аудитованою перевіркою. Метою було підвищити межі терміну служби елементів проектування до першокласних, уникаючи при цьому специфічних фреймворків для постачальників. Нейтральні зразки зберігали ідентичні алгоритми в базових та багаторівневих варіантах: синтаксичний інтерфейс та інтерфейс компілятора, який перетворює потоки токенів на абстрактні синтаксичні дерева, багаторівневий кеш з узгодженою поведінкою зчитування та блоковані числові ядра. В оцінюванні використовувалися інструментальні розподіли, підвищення, перевірені випуски та часові рамки фаз, а також парні прогони по тридцятьох незалежних початкових значеннях для порівняння інцидентів безпеки на десять мільйонів операцій, медіанного часу виконання, затримок дев'яносто п'ятого та дев'яносто дев'ятого процентилів, пропускної здатності та пікового обсягу резидентної пам'яті. Результати показали усунення витоків, подвійних звільнень, звільнень після звільнення та недійсних звільнень у межах горизонту виявлення у всіх багаторівневих варіантах, з односторонньою довірчою межею, яка встановлювала рівень інцидентів нижче 0,11 на десять мільйонів операцій. Поведінка хвостів помітно покращилася: дев'яносто п'яті процентилі зменшилися на 21,8–24,9 %, а дев'яносто дев'яті процентилі – на 22,8–27,6 % у всіх екземплярах та режимах навантаження, піковий обсяг резидентної пам'яті знизився на 10–16 %, пропускна здатність у стаціонарному стані зросла на 0,6–4,1 %, а медіанні накладні витрати часу виконання залишилися близько 1–2 %. Практично, цей підхід зменшив конкуренцію за розподільники ресурсів, дозволив цілісній програмі обмірковувати володіння та псевдоніми, а також перетворив рідкісне та дороге відновлення на передбачувану перевірку меж, пропонуючи відтворювану методологію для передового системного програмного забезпечення

Ключові слова: аудитований випуск; контроль володіння та псевдонімів; контрольні точки перевірки; контракти розподільника та дескриптора; захист псевдонімів; кодування стану типів

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 30.07.2025 Revised: 30.10.2025 Accepted: 23.12.2025

UDC 004.31: 004.056.5 DOI: 10.31649/vitce/3.2025.77

Image encryption and distribution method based on LFSR and counters

Volodymyr Luzhetskyi*

Doctor of Technical Sciences, Professor
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0000-0001-7466-7738

Mykyta Tsikhotskyi

Postgraduate Student
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0009-0005-8101-3536

Abstract. In the conditions of processing large amounts of graphic data, the task arises of developing a reliable image encryption scheme with reduced computing costs. The purpose of the study was to develop a deterministic scheme for encrypting and evenly distributing vectorised images using a shift register with linear feedback and counters. Methods of research included converting a pixel matrix to a sequence of bytes using a row-wise traversal rule, splitting the index space into equal subranges, generating pseudo-random indexes based on shift register states, and using reversible counters. The results of statistical testing demonstrate the stable characteristics of the proposed image encryption method. Encrypted test images were also evaluated for attack resistance by determining correlation coefficients between the incoming image and the encrypted one. In particular, for coloured images with a size of 512 × 512, when divided into eight subranges, the number of pixel change rate reached 99.61%, and the unified average intensity of pixel change was 32.28%, which corresponds to the upper cluster of estimates of advanced methods. The entropy of encrypted data was close to the theoretical maximum of 7.999, and the correlation between neighbouring pixels was significantly reduced and approaches zero values. Image distribution and restoration was performed without errors. The algorithm was characterised by low computational costs. The practical significance of the study consisted in ensuring reproducibility of the distribution and high cryptographic stability using mathematically simple operations, pseudo-randomness, and expanding the image encryption space to the full volume, making the proposed approach suitable for systems requiring accurate recovery and operating under limited computational resources

Keywords: secret distribution; image recovery; permutation; substitution; pseudo-random number sequence generator; image pixel correlation

Introduction

Given the current pace of information technology development, the use of digital data is growing exponentially. In accordance with the development of technologies, the requirements for protecting data that is stored, transmitted, and edited are also growing. Currently, there are many different approaches and methods that allow protecting information in the form of files of different formats. But among the file types, there is a separate category – these are images for which the use of conventional encryption methods is not appropriate, given the structure and large amount of data. Images can often contain very sensitive data, especially when they are used in critical areas of human life – medicine, military affairs, public and private secrets, etc. (Eichelberg *et al.*, 2020). Thus, contemporary science faces an important task of developing methods for protecting images in its various states, which can ensure a sufficient level of confidentiality and integrity of graphic data. In view of this, the relevance of the research is to ensure a cryptographically stable process of image encryption and uniform distribution of fragments in systems with

Suggested Citation:
Luzhetskyi, V., & Tsikhotskyi, M. (2025). Image encryption and distribution method based on LFSR and counters. *Information Technologies and Computer Engineering*, 22(3), 77-88. doi: 10.31649/vitce/3.2025.77

*Corresponding author

limited computing resources, which is especially important for Internet of Things (IoT) devices, embedded systems, and mobile solutions.

Conventional image encryption approaches based on the classic symmetric AES and DES algorithms guarantee a high level of protection, but require significant computational costs, which makes them of little use for mobile devices or IoT platforms. Therefore, there was a need for lightweight methods that can provide image encryption at minimal cost, while preserving cryptographic properties. A. Ihsan & D. Nurettin (2023) proposed a scheme that combines an affine transformation, a linear-feedback shift register (LFSR), and an XOR operation – this reduces the overhead complexity of the algorithm by using simple operations.

One of the most common encryption techniques is permutation schemes, in which information is protected by changing the order of pixels without changing their values (Babenko *et al*., 2021). Such methods require less computational resources compared to substitution or chaotic ciphers, since instead of complex nonlinear transformations, it is implemented only by replacing indexes. In most studies, this method was used in isolation: in particular, Y.-J. Sun *et al*. (2021) proposed an efficient permutation scheme based on two-dimensional logistic mapping for image encryption, which provides high entropy and resistance to statistical attacks, but has an increasing time complexity in processing large images, which limits its application in systems with strict performance requirements.

A promising area is also the use of chaotic systems for building complex multi-level encryption algorithms. Thus, Z. Liu *et al.* (2025) proposed a method for encrypting coloured images based on a discrete wavelet transform and a hyperchaotic dynamical system that combines several stages of permutation and diffusion. However, the high algorithmic complexity of the method in ($O(N \log N)$) limits the effectiveness of its application for high-resolution images.

Another area of research concerns the use of LFSR as sources of pseudorandom sequences (PRS) in image encryption schemes. LFSRs are characterised by high performance and simple hardware implementation, which makes them attractive for low-resource devices and IoT solutions. Appropriate hardware evaluations and comparisons of implementations on the Field-Programmable Gate Array (FPGA) confirm the advantages of LFSR in terms of logic element usage and performance (Dridi *et al*., 2023). In a number of applications where pseudo-random sequence (PRS) generators built using LFSR have known cryptanalysis vulnerabilities and limitations on the entropy and correlation properties of sequences compared to some chaotic generators, which requires additional measures (for example, combining multiple LFSR, nonlinear adders, or cascading with other PRS generators). Because of such trade-offs, hybrid approaches are emerging in practice – combining LFSR (as a fast hardware component) with cryptographic blocks and chaotic modules (for example, using DNA as a source of chaos and combining it with Advanced Encryption Standard (AES) components) – which increase cryptographic resistance, but significantly increase hardware and time complexity, making them less suitable for devices with limited resources. Such algorithms were presented in the papers by R. Ettiyan & V. Geetha (2023) and K. El Kinani *et al*. (2025).

An important task in encryption schemes is to divide the encrypted image into fragments so that no fragment itself contains enough information to restore the original, and reconstruction became possible only if the necessary set of parts is available. Practical implementations of block encryption often combine block permutations with chaotic diffusion, which provides flexible segmentation and localised access control, but requires the preservation of service metadata about the size and order of blocks and related access rights, which complicates exchange protocols and increases network load, one of these implementations was proposed by N. Wang *et al*. (2024). Contemporary approaches also experiment with dynamic geometric fragmentation schemes (square, L-shaped, and other geometric block extractions), which reduce the number of fragments or increase the degree of scrambling, but increase the computational complexity of the optimal splitting and recovery stage. P. Oikonomou *et al*. (2025) proposed a square image distribution method. A separate group of methods consists of approaches that use the fuzzy logic apparatus or fuzzy numbers to introduce additional uncertainty into the partitioning/distribution parameters; such methods can increase resistance to direct analysis of fragments, but do not always guarantee an even distribution of the amount of data between parts and require careful adjustment of the fuzzy rules. Y. Umadevi *et al*. (2022) proposed one of these methods of image hiding using fuzzy logic.

Thus, despite the presence of a wide range of image encryption methods, there are not enough solutions in the literature that would provide a deterministic pixel rearrangement within the entire image and simultaneously allow evenly distributing the resulting fragments without the need to save auxiliary metadata. The purpose of the study was to improve the encryption process and splitting the image into parts, with an emphasis on reducing the structural complexity of the algorithm. To achieve this goal, the following tasks were set: formalisation of the image vectorisation procedure; development of an encryption algorithm using LFSR and reversible counters; construction of a mechanism for evenly dividing the vector into n fragments and reverse image recovery. Simultaneously, the algorithm must be deterministic and meet the general requirements for evaluating the quality of encryption. The scientific originality lies in the integration of efficient LFSR components and counters to form a deterministic encryption algorithm that reduces the structural complexity of the algorithm and performs pixel rearrangement throughout the image.

Materials and Methods

The methodological component of the study was aimed at formalising and implementing an encryption algorithm and evenly distributing images into parts. It included

several stages: development of a mathematical model, software implementation, selection of a test set of images, performance evaluation based on the criteria of correct recoverability, cryptographic quality, and time complexity. The proposed image distribution algorithm contained four main methods: splitting the image into a vector; using counters and LFSR; the image (vector) encryption process itself; splitting the encrypted image into parts and the reverse recovery process.

For convenient and efficient image processing, it was proposed to perform the process of dividing pixels into a vector. Image I was defined as a matrix of pixels of size $h \times w$, where h – height, w – width. For a greyscale image, each element $I_{x,y} \in \{0, 1,..., 255\}$, where $x \in \{0,..., w\text{-}1\}$, $y \in \{0,..., h\text{-}1\}$. For a coloured image, each pixel had a triple value $I_{x,y} = (R_{x,y}, G_{x,y}, B_{x,y})$, where R – red intensity, G – green intensity, B – blue intensity, and each channel took on a value with {0, 1,..., 255}.

Transformation of matrix I in vector V was as follows: for greyscale images, the vector was $V = (V_1, V_2,..., V_N)$, where $N = h \times w$, $V_k = I_{x,y}$ for $k = y \cdot w + x$; for coloured images, there were two options for converting the image matrix I into vector V.

Option 1. Three pixel components were selected sequentially. $V = (V_1, 1, V_{1,2}, V_{1,3}, V_{2,1}, V_{2,2}, V_{2,3},..., V_{N,1}, V_{N,2}, V_{N,3})$, where $V_{k,1} = R_{x,y}$, $V_{k,2} = G_{x,y}$, $V_{k,3} = B_{x,y}$ for $k = y \cdot w + x$.

Option 2. A separate vector was formed for each pixel component, and then these vectors were combined into a single vector. $V = (V^R, V^G, V^B)$ where $V^R = (V_1^R, V_2^R,..., V_N^R)$, $V_k^R = R_{x,y}$, $V^G = (V_1^G, V_2^G,..., V_N^G)$, $V_k^G = G_{x,y}$, $V^B = (V_1^B, V_2^B,..., V_N^B)$, $V_k^B = B_{x,y}$.

These options provided for the selection of pixels in their natural display in files, namely by row-major order. The following is a formalisation of the encryption algorithm for the vector obtained after reading the image. This representation provides an unambiguous reflection of the two-dimensional image structure in a linear form, which allows correctly applying further cryptographic transformations to the vector and guarantees reproducibility of all stages of the algorithm.

Image encryption algorithm

C.E. Shannon (2001) showed that the encryption procedure can be represented as a combination of two basic transformations – permutation (diffusion) and substitution (confusion). A common approach to encrypting messages is to break them down into blocks and implement a permutation of message characters within only each block, and not within the entire message. The researchers suggest rearranging pixels not within the image blocks, but within the entire image. This increases the resistance to cryptanalysis, since the number of possible pixel permutations in the entire image is significantly greater than the number of possible permutations in the block.

As part of the development of the image encryption algorithm, two options for implementing sequential substitution and permutation were considered:

1) First substitution, then permutation (Fig. 1). Elements of the input vector are processed sequentially and, accordingly, substitution and permutation operations are performed sequentially at each step. In this approach, the permutation rule is determined by the function π:

$$\pi^{SP} = \begin{pmatrix} 0,1,...,N-1 \\ p_0,p_1,..., p_{N-1} \end{pmatrix}, \quad (1)$$

where S – substitution; P – permutation; p – element number in the encrypted vector according to the permutation rule.

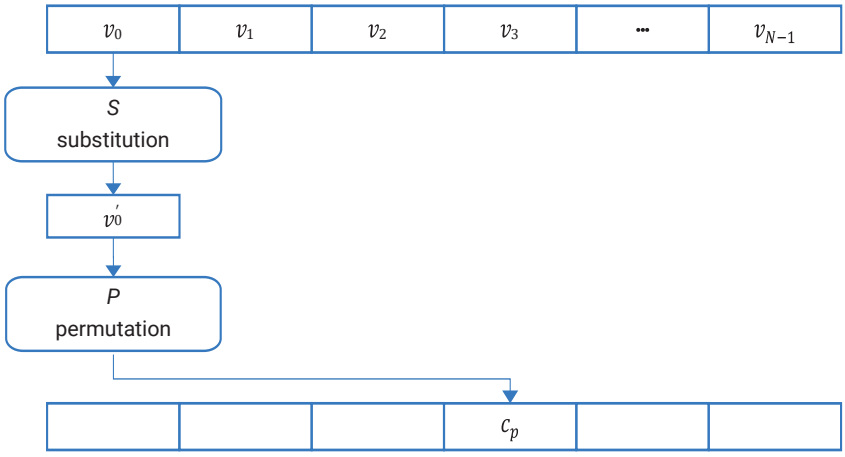


Figure 1. SP-type encryption procedur

Source: compiled by the authors

2) First the permutation, then the substitution (Fig. 2). In the second version of the encryption algorithm implementation, the input vector elements are first selected according to the permutation rule and written to the next current position (starting from zero) after applying the substitution operation. The function π has a slightly different reflection in this implementation:

$$\pi^{PS} = \begin{pmatrix} p_0,p_1,..., p_{N-1} \\ 0,1,...,N-1 \end{pmatrix}. \quad (2)$$

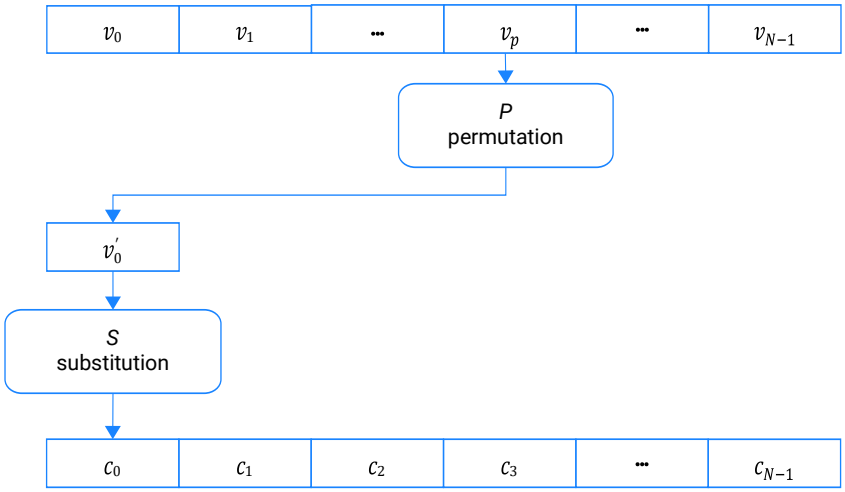


Figure 2. PS-type encryption procedure

Source: compiled by the authors

In the figures above, elements of the vector $V = \{v_0, v_1, v_2,..., v_{N-1}\}$ define elements of the image input vector; $V' = \{v'_0, v'_1, v'_2,..., v'_{N-1}\}$ – elements of the vector after performing substitution/permutation operations, $C = \{c_0, c_1, c_2,..., c_{N-1}\}$ – elements of the encrypted vector. Further, to present the proposed encryption method and its further testing, the first encryption method was chosen, namely, first performing a substitution, and then a permutation.

In this paper, it was proposed to perform substitution according to the following rule:

$$S^{SP}_{K_2}(V) = (V_i + Z(t)) \, mod \, 256;$$
$$S^{PS}_{K_2}(V') = (V'_i + Z(t)) \, mod \, 256, \quad (3)$$

where $Z(t) = (aX(t) + bY(t) + c) \bmod 256$. Here a, b, c – odd numbers that are determined by the secret key. $X(t) = (R(t) \ll 8)$ & 0xFF, where $\ll$ – means a cyclic shift to the left by 8 positions of the register status code $R(t)$. $Y(t) = (R^*(t) \ll 8)$ & 0xFF, where $R^*(t) = (R(t) \ll 8)$. This substitution acts as byte-by-byte data masking.

To implement pixel rearrangement within the entire image, it is necessary to generate pseudorandom numbers that will correspond to the new pixel position numbers p in the image. Pseudo-randomness of the sequence of numbers is ensured by using a secret key. V.A. Luzhetskyi & I.S. Horbenko (2013) proposed a generalised approach to constructing such number generators in the range from 0 to N-1. It was adapted to the conditions for solving the problem of rearranging pixels in the entire image.

Range of numbers from 0 to N-1 is split into l subranges. If $d = \frac{N}{l}$ is an integer, then each of the subranges consists of d-numbers. Each subrange is defined by the minimum and maximum values of the number:

$$D_1 = [d^1_{min}, d^1_{max}]; \quad (4)$$

If d is not an integer, then (l-1) subranges consist of $\left]\frac{N}{l}\right[$ numbers, and l-th subrange contains $N - (l-1)\left]\frac{N}{l}\right[$ numbers.

In the process of generating a pseudo-random sequence of numbers, numbers are selected from a specific subrange, the number of which is determined by the code generated by the shift register with linear feedback. The initial state of the shift register is set by the secret key K_1. The subrange number will be set r-bit binary code. With this in mind, the number of subranges is $l = 2^r$.

The selection of numbers from the subrange is carried out in a deterministic way. There are two possible options. The first is to form a sequence of numbers starting with d^1_{min}, increasing the number at each step by one. The second option is to form a sequence of numbers starting with d^1_{max}, reducing the number by one at each step. These operations are proposed to be implemented on the basis of a reversible counter CT. The counter of each subrange generates numbers from 0 to $d-1$ or from $d-1$ to 0. The availability of two options provides additional resistance to tampering, as the initial and final states can be defined separately for each counter. The initial value of the counter state is determined by the secret key component $K_2 = \{k_{2,j}\}$, where $j = 0,1,...,l-1$. Table 1 shows an example of the initial states of counters for a key $K_2 = \{0,1,1,0\}$.

Table 1. Example of the initial states of counters

K_2			
$k_{2,0}$	$k_{2,1}$	$k_{2,2}$	$k_{2,3}$
	1		0
$CT_0 := 0$	$CT_1 := d-1$	$CT_2 := d-1$	$CT_3 := 0$

Source: compiled by the authors

Changing the counter state is determined by the following rule:

$$CT_j(t+1) = CT_j(t) + (-1)^{k_2, j}, \quad (5)$$

where $CT_j(t)$ – status of j-th counter in step t.

The number selected from j-th subrange is determined based on the subrange number and the state of the corresponding counter:

$$p = j \cdot d + CT_j(t). \quad (6)$$

When using counters together with the PRS generator in the way described above, a collision may occur when in the current subrange of the vector C all numbers have already been selected in the previous steps, and the generator points to this subrange again. This situation can distort the process of encrypting/decrypting the vector by overwriting the previously selected number, which will make it impossible to restore the input image. To eliminate this, an additional mechanism for checking the current state of the counter is proposed. If $CT_j(t) = d - 1$ when $k_{2,j} = 0$ or $CT_j(t) = 0$ when $k_{2,j} = 1$, then forming the corresponding number from the subrange j does not happen. The process of forming a pseudo-random sequence of numbers is completed when the above conditions are met for all counters. Considering the specifics of implementing substitution and permutation operations using a certain secrecy, the secret key K must have four components: k_1 – initial state code of the shift register (64 bits); k_2 – a set of 8-bit parameter codes *a, b, c* for the substitution rule; k_3 – l-bit binary code that defines the initial states of counters and the rule of their operation. Thus the bit depth of the secret key K is equal to $(88 + l)$ bit.

Operations that implement substitution and permutation rules to perform the image encryption process have been described and defined above. This process is now presented as a single aggregate function to further simplify the description of image distribution, recovery, and decryption processes. Let the encryption for the *SP* method be described as:

$$E_K^{SP}(\mathrm{V}) = S_{K_2}^{SP}(\mathrm{V}) \circ P_{K_1, K_3}(\mathrm{V}'), \quad (7)$$

where S_{K_2} – substitution function; P_{K_1, K_3} – permutation function.

Similarly, the encryption function is presented for the *PS* method:

$$E_K^{PS}(\mathrm{V}) = P_{K_1, K_3}(\mathrm{V}) \circ S_{K_2}^{PS}(\mathrm{V}'). \quad (8)$$

Thus, the authors proposed and described a new image encryption method that combines substitution and controlled permutation operations using a pseudorandom sequence generator based on a linear feedback shift register (LFSR). After encrypting the input image, vector $C = (c_0, c_1, \ldots, c_{N-1})$ is divided into n subvectors for the purpose of organising distributed storage without the ability to restore the full image if only a part of the subvectors is present. For this purpose, a secret distribution scheme of the type (n, n), which guarantees that only if all n parts are possible to restore the original content correctly, which increases the level of information security.

For subvector distribution, it is proposed to use a cyclic uniform distribution, which ensures uniform data filling of each subvector and allows maintaining a linear correspondence between the elements of the original and encrypted vector, which simplifies the decryption process and minimises overhead calculations. This approach, in comparison with chaotic or random distribution schemes, is quite stable for structural analysis, but simultaneously remains controlled and deterministic, which is critical for further image restoration.

Each subvector of distributed image – $P_y = \{p_{y,0}, p_{y,1}, p_{y,2}, \ldots\}$, where y {0, 1,..., $n-1$} , is formed from the elements of the vector C for which the condition is met:

$$p_{y,q} = c_i, where\ y = i \bmod n, q = y + \left\lfloor \frac{i}{n} \right\rfloor \cdot n. \quad (9)$$

Thus, a sequence of subvectors is formed $\{P_0, P_1, \ldots, P_{n-1}\}$, the dimensions of which do not exceed $\frac{N}{n} \pm 1$. To restore the input image, a complete set of $\{P_0, P_1, \ldots, P_{n-1}\}$ parts is required.

Since the cyclic uniform approach was used for the distribution, the indexes of the elements of the encrypted vector were assigned in a deterministic order, which preserves a one-to-one correspondence between the elements of each part P_y and their positions in C. This allows generating a reverse transformation that restores C by simply combining fragments with fixed indexes according to the original distribution order.

For each part $P_y \subset C$, restoring an element of vector C occurs using the equation:

$$c_i = p_{y,q}\ , where\ y = i \bmod n;\ q = y + \left\lfloor \frac{i}{n} \right\rfloor \cdot n. \quad (10)$$

After assembling the parts from encrypted vector C to restore the image, sequential substitution and permutation operations are performed, the order of which depends on the original encryption method – *SP* or *PS*. Since the substitution operation S_{K_2} is bijective, its inverse $S_{K_2}^{-1}$ restores information before performing a substitution. Accordingly, the inverse substitution operation $S_{K_2}^{-1}$ is performed as follows:

$$S_{K_2}^{-1, SP}(V') = (V_i' - Z(t)) \ mod\ 256; \quad (11)$$

$$S_{K_2}^{-1, PS}(C) = (C_i - Z(t)) \ mod\ 256. \quad (12)$$

Below is a mathematical representation of the decryption function $D_K(C)$, as a result of which the initial vector V is obtained:

$$D_K^{SP}(\mathrm{C}) = P_{K_1, K_3}(\mathrm{C}) \circ S_{K_2}^{-1, SP}(V'); \quad (13)$$

$$\mathrm{D}_K^{PS}(\mathrm{C}) = S_{K_2}^{-1}, PS\,(\mathrm{C}) \circ P_{K_1, K_3}(V'). \quad (14)$$

The last step is to restore the image from the resulting vector by folding it back into pixels and presenting it as an image file. This step involves accurately reproducing the spatial structure of the image based on the stored index order generated during encryption. As a result of folding the provided parts and decrypting the image, the initial secret image is formed.

Results and Discussion

During the experiments, the cryptographic characteristics of the proposed algorithm were evaluated on test sets for greyscale and coloured images with a size of 512 × 512 pixels (.tiff format). The following secret key parameters are selected for testing: K_1 = 0xA1B2C3D4E5F60708, K_2 = {5, 7, 11}, K_3 = {0, 1, 0, 1} for 4 subranges and K_3 = {0, 1, 0, 1, 1, 0, 0, 1} for 8 subranges. The feedback positions in the shift register are defined as [63, 62, 60, 59]. For this purpose, statistical testing of the encryption method was performed using the National Institute of Standards and Technology (NIST SP 800-22) statistical test kit (Rukhin *et al.*, 2010). Table 2 shows the results of statistical tests that were performed on a 2 Mbit encrypted data sequence. To ensure sufficient data sampling, 12 greyscale images were encrypted and sequentially written to the test file as bits, and up to 4 test images were used for coloured images.

Table 2. Results of statistical testing using NIST SP 800-22

Coloured 4		Coloured 8		Greyscale 4		Greyscale 8		Statistical test
P	Fraction	P	Fraction	P	Fraction	P	Fraction	
0.534	9/10	0.739	10/10	0.534	10/10	0.122	10/10	Frequency
0.534	10/10	0.35	10/10	0.739	9/10	0.739	10/10	BlockFrequency
0.442	9/10	0.545	10/10	0.545	10/10	0.094	10/10	CumulativeSums
0.534	10/10	0.739	10/10	0.534	10/10	0.122	10/10	Runs
0.213	10/10	0.911	9/10	0.122	10/10	0.534	10/10	LongestRun
0.213	10/10	0.213	10/10	0.534	10/10	0.739	10/10	Rank
0.35	10/10	0.911	10/10	0.35	10/10	0.739	10/10	FFT
0.35	10/10	0.35	10/10	0.008	10/10	0.911	10/10	Overlapping template
0.066	10/10	0.213	9/10	0.350	10/10	0.35	9/10	Universal
0.122	10/10	0.911	10/10	0.122	10/10	0.911	10/10	ApproxEntropy
0.739	10/10	0.534	10/10	0.534	10/10	0.739	10/10	LinearComplexity
–	10/10	–	10/10	–	10/10	–	10/10	Serial median
–	9.91/10	–	9.89/10	–	9.86/10	–	9.91/10	NonOverlapping template median
–	5/5	–	3.88/4	–	7/7	–	6/6	RandomExcursions median
–	5/5	–	3.89/4	–	7/7	–	6/6	RandomExcursions variant median

Source: compiled by the authors

The obtained results of statistical testing demonstrate the stable characteristics of the proposed image encryption method. For most tests, P values exceed the set significance threshold $\alpha = 0.01$, which indicates that there are no detected signs of statistical non-randomness in the generated bit sequences. This indicates that the method provides a sufficient level of entropy and uniformity of distribution, which are key criteria for cryptographically stable transformations. For coloured images with an increased number of subranges, there is a slight increase in the uniformity of bit distribution and a more uniform distribution of P over the intervals. This indicates improved diffusion as the number of subranges increases, which reduces the likelihood of local structures appearing in the encrypted data. In greyscale images with fewer subranges, the algorithm shows only slightly lower performance for individual tests (for example, in the case of frequency and block frequency tests), but these values still remain within the acceptable range ($P \geqslant \alpha$). The Fraction value indicates the number of successfully completed tests. According to the results, almost all tests were passed successfully, a deviation of 1 unsuccessful test is acceptable. Figure 3 shows a histogram of the distribution of coefficients depending on the image type and the specified number of subranges.

Figure 4 shows the pixel brightness distribution (in the range of values from 0 to 255) for the input greyscale image. The nature of the curve indicates the presence of pronounced peaks and troughs, which reflects the predominance of certain brightness levels. Such unevenness is typical for natural or visually understandable images, since their structure and content form statistical patterns that can be used by an attacker to simplify cryptanalysis. Figure 5 shows the distribution for the encrypted image, which is almost uniform, which indicates a high level of randomness of the received data, which significantly complicates their cryptanalysis and is a sign of a stable encryption algorithm.

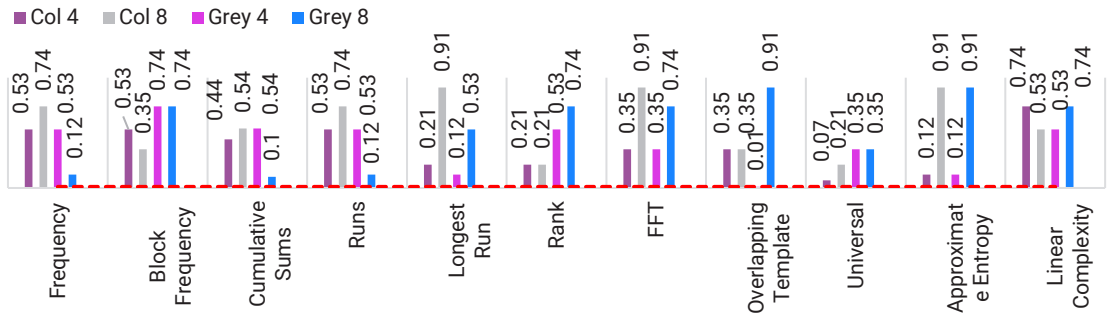


Figure 3. Histogram of NIST statistical testing results
Note: red line indicates the permissible threshold value $\alpha = 0.01$
Source: compiled by the authors

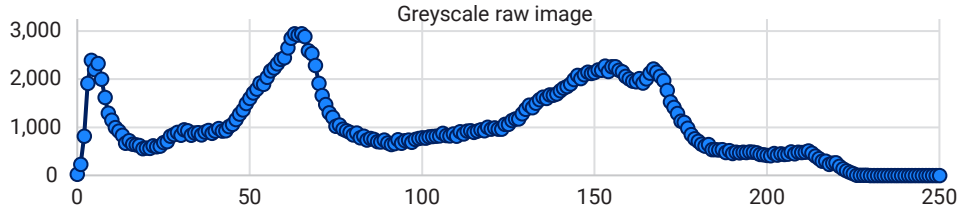


Figure 4. Pixel brightness distribution in the input greyscale image
Source: compiled by the authors

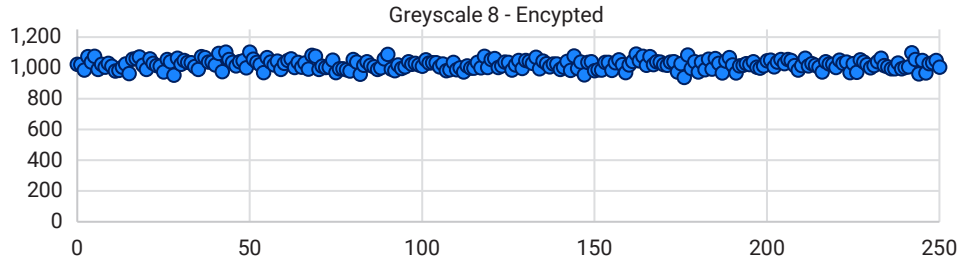
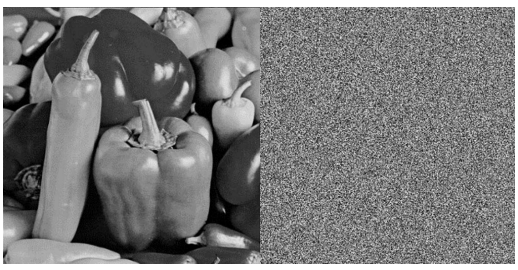


Figure 5. Pixel brightness distribution in the encrypted greyscale image
Source: compiled by the authors

Figure 6 shows the results of performing an encryption scheme, where the input image has a natural structure and easily recognisable content, and the encrypted image visually resembles random noise without any noticeable dependencies. This type of encrypted image is typical for systems with efficient diffusion, where each change in input pixels significantly affects the output result. This indicates that the algorithm is reliably resistant to attacks based on visual analysis.



A B

Figure 6. Encryption results for the greyscale Peppers test image
Note: A – input image; B – encrypted image
Source: compiled by the authors

For coloured images, the brightness distribution was analysed separately for each of the three colour channels (*R, G, B*), which allows identifying specific features of their structure. As can be seen from Figure 7, the input image has pronounced peak values in each channel, reflecting the uneven colour distribution and the presence

of dominant shades. The encryption result (Fig. 8) demonstrates that all channels have a uniform distribution, which indicates an effective destruction of the initial correlations between pixels in each of the channels and makes it impossible to restore input characteristics based on statistical analysis.

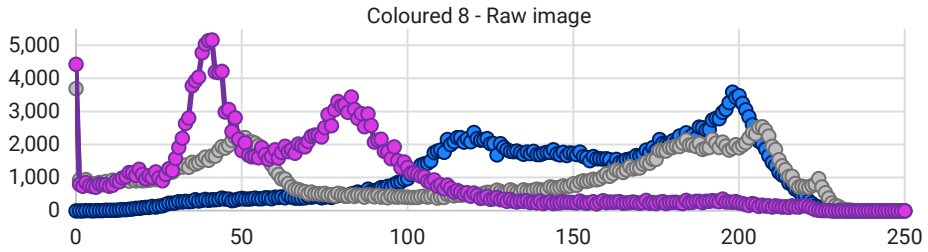


Figure 7. Pixel brightness distribution in the input coloured image
Source: compiled by the authors

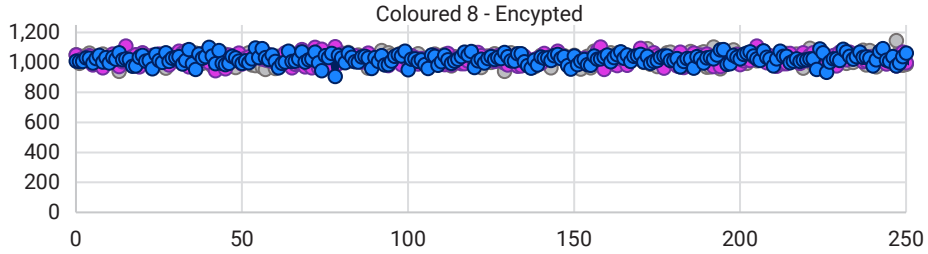


Figure 8. Pixel brightness distribution in the encrypted coloured image
Source: compiled by the authors

Similar to the greyscale example, encryption turns the coloured input image into noise and makes it impossible to recognise the secret (Fig. 9). This visual transformation confirms the efficiency of diffusion in the algorithm, which is a critical factor for resistance to visual analysis. It also demonstrates the absence of a residual structure that could have been used to reconstruct the original image.

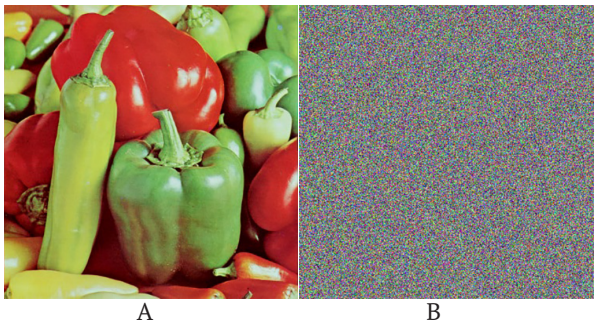


Figure 9. Encryption results for the coloured Peppers test image
Note: A – original image; B – encrypted image
Source: compiled by the authors

Since the specifics of image encryption differ significantly from the process of encrypting ordinary bit data (text), methods for evaluating cryptographic stability also require some adaptation. Performing statistical testing inherent in bit sequence encryption does not always fully reflect the key characteristics of encrypted images, in particular, their resistance to attacks. Therefore, in addition to NIST tests, the correlation between neighbouring pixels is additionally analysed, which is described by the NPCR (Number of Pixels Change Rate) and UACI (Unified Average Changing Intensity) parameters. The NPCR indicator reflects the percentage of pixels that changed their value with a slight change in the input data, and is used to estimate the sensitivity of the algorithm to the initial conditions: values close to 100% indicate high stability, when even a change in one pixel in the original image leads to a completely unpredictable result. The UACI parameter determines the average intensity of pixel brightness changes between two images and reflects the degree of destruction of the initial correlations; according to

Y. Wu *et al*. (2011) and Y. Alghamdi & M. Arslan (2024), its optimal value for stable algorithms is in the range of 32-34%. In addition, the entropy of the initial and encrypted images was calculated as a measure of the randomness of the pixel distribution, which allows estimating the approximation of the result to a completely random one. The correlation requirements between pixels determine: the closer this value is to zero, the smaller the relationship. The parameters of cryptographic stability and pixel correlation are shown in Table 3.

Table 3. Parameters of pixel correlation and NPCR, UACI

		Coloured 4	Coloured 8	Greyscale 4	Greyscale 8
Entropy	original	7.6698	7.6698	7.5715	7.5715
	encrypted	7.9997	7.9997	7.9993	7.9993
Horizontal correlation	original	0.9704	0.9704	0.9792	0.9792
	encrypted	0.0004	0.0008	-0.0016	-0.0003
Vertical correlation	original	0.9715	0.9715	0.9826	0.9826
	encrypted	-0.0014	-0.0009	0.0006	0.0006
Diagonal correlation	original	0.9576	0.9576	0.9680	0.9680
	encrypted	0.0009	0.0006	-0.0002	0.0040
NPCR		99.61%	99.61%	99.61%	99.60%
UACI		32.18%	32.28%	31.00%	31.00%

Source: compiled by the authors

In the table below, the results are presented for four options for using the encryption scheme: coloured and greyscale images with 4 and 8 subranges, respectively. For coloured images, the entropy of encrypted data is close to the maximum theoretical value (7.9997), which indicates a high level of randomness, and the UACI of ~32.2% corresponds to the optimal range for protection against attacks that analyse local changes and dependencies between pixels of the input image and the encrypted one. The NPCR value consistently exceeds 99.6%, which demonstrates the high sensitivity of the algorithm to changing even one pixel. For greyscale, the entropy of encrypted data also shows results close to 8 (7.9993). UACI (~31%) also meets the sustainability requirements, while NPCR retains a value of approximately 99.6%. Comparative analysis shows that increasing the number of subranges from 4 to 8 has a positive effect on increasing the degree of randomness of the encrypted image. It is also noted that the method shows better results when working with colour images due to an increase in the natural difference in brightness distribution and colour depth compared to greyscale. Table 4 shows estimates of the NPCR and UACI values of the proposed encryption algorithm and other known encryption algorithms. For evaluation, a 512×512 coloured image was taken divided into 8 subranges, the encryption results of which showed the best results for the proposed method: NPCR = 99.61%, UACI = 32.28%.

Table 4. Comparison of NPCR and UACI (512 × 512 coloured images)

No.	Author	NPCR (%)	UACI (%)
1	Z. Liang *et al.* (2021)	99.6	33.3
2	X. Wang *et al*. (2019)	99.6	31.5
3	Y. Abanda & A. Tiedeu (2016)	99.6	32.05
4	B. Stoyanov & K. Kordov (2015)	99.5-99.7	31-32
5	Y. Alghamdi *et al*. (2022)	99.5-99.7	31-32
6	Authors' scheme	99.61	32.28

Source: compiled by the authors

Z. Liang *et al*. (2021) proposed a scheme based on a five-dimensional chaotic system using DNA coding and genetic operations. This study demonstrated NPCR and UACI at a fairly high level, which is explained by multi-level diffusion at the bit operation level and the use of chaos. Compared to the algorithm under study, this approach has a higher implementation complexity and a larger number of parameters to configure, which makes it difficult to use on platforms with limited resources. Such a scheme shows better diffusion results, the disadvantage is increased implementation complexity and higher computational costs.

X. Wang *et al*. (2019) proposed an encryption algorithm that uses an S-box formed from a chaotic sequence. This results in increased substitution nonlinearity and stable NPCR ≈ 99.6, UACI ≈ 31 – 32. The main difference from the presented algorithm is the use of nonlinear substitution. The algorithm provides sufficient statistical stability, but creating and verifying an S-box increases complexity and complicates hardware implementation.

Simple but reliable combinations of chaotic maps and "mixing" to generate permutations and diffusion are used in the study by Y. Abanda & A. Tiedeu (2016). The results show UACI ≈ 32.05% and NPCR ≈ 99.6. Such an algorithm provides stable results with minimal implementation complexity, but provides less control over determining the sequence of transformations and complicates formal proof of security.

B. Stoyanov & K. Kordov (2015) presented an encryption scheme using polynomial chaotic mappings, which showed UACI results in the range of 31-32% with a sufficient NPCR value. The difference is a mathematically more complex model of the PRS generator, which gives a wider key space. In this scheme, balanced statistical characteristics are noted at the same level as the author's, but it is worth considering the increase in hardware complexity of implementation.

Y. Alghamdi *et al*. (2022) considered algorithms for devices with limited resources, where local block permutations and simple chaotic maps are used. Test statistics showed NPCR ≈ 99.5 – 99.7 and UACI ≈ 31 – 32. This indicates that the algorithm is better adapted to hardware limitations compared to previous methods, but the correlation indicators are less than the required minimum and, accordingly, the author's scheme.

The obtained NPCR and UACI values showed that the proposed algorithm belongs to the group of advanced efficient solutions. NPCR = 99.61% indicates a high sensitivity of the encryption to minimal changes in the input image. UACI = 32.28% shows a good intensity of change, indicating effective diffusion, although this figure is slightly lower than in some examples with deeper multilevel diffusion (UACI ≳ 33%, which is the reference value). The results obtained indicate a balance between cryptographic stability and computational ease of implementation. This confirms the feasibility of using the algorithm in systems with strict resource constraints and the need for guaranteed and deterministic image recovery.

Conclusions

This paper proposed a deterministic scheme for encryption and uniform distribution of vectorised images based on a linear feedback shift register and controlled reversible counters. The main objectives – to provide linear time computational complexity, deterministic and reproducible partitioning without storing additional metadata, and to achieve a cryptographically acceptable level of randomness of encrypted data – were successfully achieved and experimentally validated. The proposed algorithm demonstrated linear complexity with respect to the number of pixels (O(N)), since each iteration requires one single invocation of the PRS generator and a constant set of counter update and write operations. The cyclic uniform distribution mechanism ensured uniform filling of subvectors and allowed restoring the original vector only if all n parts are available, thereby satisfying the requirements of an (n, n) secret distribution scheme. During reconstruction, zero reconstruction errors and no pixel degradation were observed when combining fragments and performing decryption.

The results of testing according to NIST SP 800-22 showed no pronounced signs of statistical non-randomness in the generated bit sequences (p values are greater than the threshold $\alpha = 0.01$ for most tests and the proportion of successfully passed tests met the required criteria). The entropy of the encrypted images approached the theoretical maximum of 7.999, which indicates a uniform distribution of pixel intensities. Correlation coefficients between adjacent pixels horizontally, vertically, and diagonally dropped from high values of 0.95-0.98 in the original images to values close to zero or small negative values in encrypted images (about 10^{-3}-10^{-4}), which confirmed the destruction of spatial dependencies. Attack resistance indicators based on the analysis of differences between open and encrypted images – NPCR > 99.6% and UACI 31-32% – met the generally accepted criteria for effective avalanche behaviour and sufficient diffusion between original-ciphertext pairs. Due to the simplicity of operations (bit shifts, modular arithmetic, such as addition with a modulus of 256) and minimal memory requirements, the proposed scheme is well suited for implementation on embedded platforms, mobile devices, and other systems with limited computational resources. The absence of the need to store metadata for fragment recovery reduces network and memory overhead in distributed storage.

Since some practices use a cryptographic threshold (k, n) a secret distribution scheme, so the prospects for further research are to develop a scheme that will allow effective application of the recovery mechanism to images encrypted by the proposed method. Future research should focus on optimising the implementation of the algorithm, as this will reduce encryption time, which is especially important for devices with limited resources or systems that transmit data in real time. It is also advisable to strengthen the diffusion properties in order to bring the UACI indicators closer to the reference level (~ 33%), while not complicating the overall structure of the algorithm, which will ensure a balance between safety and efficiency. In addition, advanced testing of the algorithm on various sets of images of different formats will allow for a deeper study of its resistance to statistical attacks, identify possible relationships between data types and encryption efficiency, and identify potential limitations or weaknesses of the proposed approach.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Abanda, Y., & Tiedeu, A. (2016). Image encryption by chaos mixing. *IET Image Processing*, 10(10), 742-750. doi: 10.1049/iet-ipr.2015.0244.

[2] Alghamdi, Y., & Arslan, M. (2024). Image encryption algorithms: A survey of design and evaluation metrics. *Journal of Cybersecurity and Privacy*, 4(1), 126-152. doi: 10.3390/jcp4010007.

[3] Alghamdi, Y., Munir, A., & Ahmad, J. (2022). A lightweight image encryption algorithm based on chaotic map and random substitution. *Entropy*, 24(10), article number 1344. doi: 10.3390/e24101344.

[4] Babenko, V., Myroniuk, T., & Krivous, H. (2021). Algorithms for application of permutation operations controlled by information for implementation of cryptographic transformation of information. *Bulletin of Cherkasy State Technological University*, 26(3), 44-58. doi: 10.24025/2306-4412.3.2021.247252.

[5] Dridi, F., El Assad, S., Wajih, E., & Machhout, M. (2023). Design, hardware implementation on FPGA and performance analysis of three chaos-based stream ciphers. *Fractal and Fractional*, 7(2), article number 197. doi: 10.3390/fractalfract7020197.

[6] Eichelberg, M., Kleber, K., & Kämmerer, M. (2020). Cybersecurity in PACS and medical imaging: An overview. *Journal of Digital Imaging*, 33(6), 1527-1542. doi: 10.1007/s10278-020-00393-3.

[7] El Kinani, K., Amounas, F., Bendaoud, S., & Bayane, Y. (2025). Hybrid approach for IoT-based medical image encryption and compression using modified AES and chaos theory. In Y. Farhaoui, T. Herawan, A.L. Imoize & A.E. Allaoui (Eds.), *Intersection of artificial intelligence, data science, and cutting-edge technologies: From concepts to applications in smart environment. ICAISE 2024. Lecture notes in networks and systems* (Vol. 1353, pp. 390-395). Cham: Springer. doi: 10.1007/978-3-031-88304-0_54.

[8] Ettiyan, R., & Geetha, V. (2023). A hybrid logistic DNA-based encryption system for securing the Internet of Things patient monitoring systems. *Healthcare Analytics*, 3, article number 100149. doi: 10.1016/j.health.2023.100149.

[9] Ihsan, A., & Nurettin, D. (2023). Improved affine encryption algorithm for color images using LFSR and XOR encryption. *Multimedia Tools and Applications*, 82(5), 7621-7637. doi: 10.1007/s11042-022-13727-w.

[10] Liang, Z., Qiuxia, Q., Changjun, Z., Ning, W., Yi, X., & Wenshu, Z. (2021). Medical image encryption algorithm based on a new five-dimensional three-leaf chaotic system and genetic operation. *PLOS One*, 16(11), article number e0260014. doi: 10.1371/journal.pone.0260014.

[11] Liu, Z., Li, C., Zhang, C., & Yang, X. (2025). Dual-domain image encryption scheme based on fractional wavelet transform and hyperchaotic system. *Physica Scripta,* 100(3), article number 035234. doi: 10.1088/1402-4896/adb529.

[12] Luzhetskyi, V.A., & Horbenko, I.S. (2013). Method for forming permutations of an arbitrary number of elements. *Information Security*, 15(3), 262-267.

[13] Oikonomou, P., Kranas, G.K., Sapounaki, M., Spathoulas, G., Aretaki, A., Kakarountas, A., & Adam, M. (2025). Square-based division scheme for image encryption using generalized Fibonacci matrices. *Mathematics*, 13(11), article number 1781. doi: 10.3390/math13111781.

[14] Rukhin, A., *et al*. (2010). *A statistical test suite for random and pseudorandom number generators for cryptographic applications (NIST SP 800-22 Rev. 1)*. doi: 10.6028/NIST.SP.800-22r1a.

[15] Shannon, C.E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1), 3-55. doi: 10.1145/584091.584093.

[16] Stoyanov, B., & Kordov, K. (2015). Image encryption using Chebyshev map and rotation equation. *Entropy*, 17(4), 2117-2139. doi: 10.3390/e17042117.

[17] Sun, Y.-J., Zhang, H., Wang, X.-Y., & Wang, M.-X. (2021). Bit-level color image encryption algorithm based on coarse-grained logistic map and fractional chaos. *Multimedia Tools and Applications*, 80, 12155-12173. doi: 10.1007/s11042-020-10373-y.

[18] Umadevi, Y., Ashwini, M., & Savitha, N. (2022). Fuzzy logic-based parallel data embedding technique for image steganography. *International Journal of Innovative Research in Computer and Communication Engineering*, 10(8). doi: 10.15680/IJIRCCE.2022.1008045.

[19] Wang, N., Wang, X., Liu, A., Wang, W., Ding, Y., Wu, X., & Du, X. (2024). An image partition security-sharing mechanism based on blockchain and chaotic encryption. *PLOS One*, 19(7), article number e0307686. doi: 10.1371/journal.pone.0307686.

[20] Wang, X., Çavuşoğlu, Ü., Kaçar, S., Akgül, A., Pham, V.-T., Jafari, S., Alsaadi, F.E., & Nguyen, X. Q. (2019). S-box based image encryption application using a chaotic system without equilibrium. *Applied Sciences*, 9(4), article number 781. doi: 10.3390/app9040781.

[21] Wu, Y., Noonan, J., & Agaian, S. (2011). NPCR and UACI randomness tests for image encryption. *Journal of Selected Areas in Telecommunications*, 2011, 31-38.

Метод шифрування та розподілу зображень на основі LFSR та лічильників

Володимир Лужецький

Доктор технічних наук, професор
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0001-7466-7738

Микита Ціхоцький

Асистент
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0005-8101-3536

Анотація. У сучасних умовах обробки великих обсягів графічних даних постає завдання розробки надійної схеми шифрування зображень зі зменшенням обчислювальних витрат. Метою дослідження було розробити детерміновану схему шифрування та рівномірного розподілу векторизованих зображень із використанням регістра зсуву з лінійним зворотним зв'язком і лічильників. Методи роботи включали перетворення матриці пікселів у послідовність байтів за правилом обходу по рядках, розбиття індексного простору на рівні піддіапазони, генерацію псевдовипадкових індексів на основі станів регістра зсуву та використання реверсивних лічильників. Результати статистичного тестування демонструють стійкі характеристики запропонованого методу шифрування зображень. Також було проведено оцінку зашифрованих тестових зображень до стійкості атаки шляхом визначення коефіцієнтів кореляції між вхідним зображенням та зашифрованим. Зокрема, для кольорових зображень розміром 512×512 при розбитті на вісім піддіапазонів коефіцієнт зміни кількості пікселів склав 99,61 %, а уніфікована середня інтенсивність зміни пікселів – 32,28 %, що відповідає верхньому кластеру оцінок сучасних методів. Ентропія зашифрованих даних наближена до теоретичного максимуму та склала 7,999, а кореляція між сусідніми пікселями істотно зменшена і наближається до нульових значень. Розподіл та відновлення зображення виконується без похибок. Алгоритм відзначається низькими обчислювальними витратами. Практична цінність дослідження полягає в забезпеченні відтворюваності розподілу й високу криптографічну стійкість з використанням математично простих операцій, псевдовипадковості та розширення простору шифрування зображення до повного обсягу, що робить підхід придатним для систем із вимогою точного відновлення й працюють з обмеженими обчислювальними ресурсами

Ключові слова: розподіл секрету; відновлення зображення; перестановка; підстановка; генератор псевдовипадкової послідовності чисел; кореляція пікселів зображення

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 11.09.2025 Revised: 20.11.2025 Accepted: 23.12.2025

UDC 004.03 DOI: 10.31649/vitce/3.2025.89

Application of chaos theory to improve resilience of encryption systems in information technology

Volodymyr Lukhanin*

PhD in Physical and Mathematical Sciences, Assistant
Kharkiv National University of Radio Electronics
61166, 14 Nauky Ave., Kharkiv, Ukraine
https://orcid.org/0000-0003-4328-929X

Abstract. The study aimed to provide a theoretical justification for the use of chaotic dynamical systems to enhance the strength of cryptographic keys. The research methodology was based on theoretical, comparative and critical analysis of scientific sources to assess the potential of chaotic systems. The study determined that chaotic maps provide high entropy, long periods and unpredictability of the generated sequences due to their sensitivity to initial conditions, which is confirmed by the Shannon entropy calculations and positive Lyapunov exponents. The use of hash functions and mechanisms for updating the internal state eliminated statistical correlations and increased the resistance of generators to cryptanalysis. The study demonstrated that the sequences obtained on the basis of the logistic mapping and the Lorentz system pass the standard statistical tests of NIST SP 800-22, demonstrating uniformity of distribution and absence of correlations. The use of the Chua circle as an analogue circuit provides physically implemented True Random Number Generators with low power consumption, suitable for resource-limited Internet of Things systems. The scheme with the integration of several chaotic maps has proven to increase the key space and increase the resistance to statistical attacks compared to traditional PseudoRandom Number Generators. The study determined that chaotic generators are able to provide forward and backward secrecy by updating the internal state of the system, which prevents the sequences from repeating. Chaotic generators have advantages over traditional PseudoRandom Number Generators due to their very long periods and sensitivity to initial conditions, but their effectiveness depends on cryptographic post-processing and the correct choice of parameters. The study recommended the use chaotic systems as an additional source of entropy in software and hardware implementations, in particular, in lightweight cryptographic solutions for the Internet of Things, sensor networks and mobile devices. The practical significance is determined by the application of the results by developers for secure encryption, researchers for random number generation, and Internet of Things engineers for device security

Keywords: nonlinear dynamics; random number generators; cryptographic entropy; chaotic attractors; initialisation vectors; topological transitivity; cryptographic extraction

Introduction

Ensuring the resilience of cryptographic systems is one of the key challenges of information security in the 21st century. The growth in data transmission, the proliferation of cloud services and the rapid development of the Internet of Things (IoT) create new risks to the confidentiality and integrity of information. Traditional encryption methods, including symmetric and asymmetric algorithms, have proven to be effective, but their sustainability is gradually being questioned due to the increase in computing power and the emergence of quantum technologies (Fernández-Caramès & Fraga-Lamas, 2020). This creates a need to find new approaches to cryptographic key generation that would provide a significant level of unpredictability and security. One of these promising areas is the use of chaotic dynamical systems capable of generating sequences with high entropy and complex structure.

The scientific discourse further addresses the use of chaos in cryptography. M. Ali *et al.* (2025) proposed an approach to building an encryption system using geometric permutations and dynamic substitutions. The authors showed that the combination of chaotic maps with new methods of data structuring can significantly increase the

Suggested Citation:
Lukhanin, V. (2025). Application of chaos theory to improve resilience of encryption systems in information technology. *Information Technologies and Computer Engineering*, 22(3), 89-100. doi: 10.31649/vitce/3.2025.89

*Corresponding author

cipher's resistance to linear and differential analysis. This forms a new combination strategy that improves the resistance of classical chaotic algorithms. The practical application of chaotic models in resource-constrained environments was demonstrated by T.A. Dhopavkar *et al.* (2022). The study used Tinkerbell and Duffing maps to create a data protection scheme for IoT systems. The results proved that chaotic maps can provide both lightweight algorithms and a high level of security even in devices with limited computing capabilities. This proved the suitability of chaotic maps for IoT applications with low resource requirements.

A. Belazi *et al.* (2022) addressed the protection of medical images, where data quality and reliability are critical. The study improved on the sinus tangent map and demonstrated that it produces uniform and statistically stable sequences that guarantee high ciphertext entropy. This meant that chaos proved to be an effective tool in medical cryptography. In the field of satellite imagery, promising results were shown by A. Kumar & M. Dua (2021). The study proposed to use the cosine transform in combination with chaotic maps, which improved the quality of key generation and provided an additional level of data protection. This proved the versatility of chaos as a tool for cryptographic applications in various industries.

The tendency to combine different chaotic models was reflected by M. Kumar & D. Ch (2025). The study demonstrated that merging chaotic maps with multilevel mixing techniques makes it possible to achieve system robustness to statistical analysis. In particular, multi-level shuffling significantly complicates the ability to predict keys, making brute-force attacks almost impossible. This formed an approach to integrating chaotic maps with multilevel transformations, which increases the cryptographic strength of systems by complicating the key structure. A similar direction was pursued in E. Faure *et al*. (2024), where authors proposed enhancements to classical chaos-based encryption schemes to improve key agreement and data security. An additional perspective was revealed by J. Jackson & R. Perumal (2025), employing fractional-order chaotic maps. The study determined that the use of more complex mathematical models can generate sequences with increased unpredictability and a higher degree of security. This extended the capabilities of traditional models and increased the level of cryptographic security.

In the Ukrainian scientific space, there is also considerable interest in the topic of cryptography and chaotic models. The study by A. Shandyba (2025) contributed to the practical application of chaos for information security, in particular in the field of digital watermarks. The study demonstrated that the use of chaotic maps when embedding markers ensures the system's resistance to attacks and preserves the authenticity of multimedia data. This confirmed the potential of chaos as a tool not only for key generation but also for expanding the range of cryptographic applications. O. Krulikovskyi *et al*. (2024) analysed the periodicity of time series generated by a logistics map, taking into account the limited accuracy of digital computing. The study highlighted that the limitations of machine arithmetic can lead to the degradation of chaos and the emergence of periodic structures, which directly affects the reliability of cryptographic generators based on chaotic models. The study revealed critical aspects of the implementation of chaotic algorithms in digital systems and emphasised the need to incorporate computational accuracy when developing chaos-based cryptographic mechanisms.

Despite the numerous results, most studies focus on applied tasks, image security, IoT, or medical data. However, there is a lack of generalised theoretical models that systematically describe how chaos can be used to generate keys in a broader cryptographic context. In addition, much of the research is focused on individual chaotic maps, which could not be used to assess the potential of a comprehensive approach to combining them. This creates a gap between the theoretical basis and practical applications. Therefore, the study aimed to theoretically study the possibilities of using chaotic dynamic systems to increase the strength of cryptographic keys. To achieve this goal, the following tasks were performed: to systematise and formalise the properties of chaotic dynamic systems that determine their suitability for cryptographic applications, to present a three-stage model of chaos-based key generation, to evaluate the efficiency, practical stability and potential of the model for use in symmetric, asymmetric and hybrid cryptosystems.

Materials and Methods

The study included a theoretical comprehensive review of chaotic dynamical systems and their models, the creation of an abstract key generation scheme, an assessment of practical applications, limitations, and a comparison with traditional generators to determine their cryptographic potential. The research included four stages. At the first stage, the method of generalising scientific approached to chaotic dynamical systems and their application in cryptography was applied in the following areas: symmetric and asymmetric algorithms, steganography and multimedia, hardware solutions (True Random Number Generator (TRNG), IoT). This systemised and assessed the level of entropy, unpredictability, and resistance to cryptanalysis. The task of this stage was to determine the potential of chaotic systems as an effective source of entropy for encryption and information security protocols.

At the second stage, the method of theoretical analysis was used to test the suitability of chaotic systems for generating random sequences in cryptography. The mathematical and physical models of chaos, such as the logistic mapping (provides a simple implementation and is used to generate pseudorandom numbers), the Lorentz system (demonstrates complex multidimensional trajectories with high unpredictability), and the Chua circle (can be used for chaos at the hardware level, making it suitable for TRNG and IoT solutions), were considered to test their suitability for generating random sequences in cryptography. The choice of these models was justified by the fact that they

represent different levels of complexity, from simple mathematical constructions to multidimensional and hardware solutions. The study analysed their properties (topological mixability, transitivity, high entropy, long periods, and practical non-repeatability) and presented a generalised abstract model of key generation, which includes quantisation of values, cryptographic extraction (SHAKE256, Hash-based Message Authentication Code (HMAC)-based Key Derivation Function (HKDF)), and internal state update (reseeding). The stage aimed to show how chaotic trajectories can be transformed into cryptographically strong key material with high entropy and unpredictability.

The third stage presented an abstract model of key generation that combines the dynamics of chaotic systems with cryptographic primitives to obtain a stable key material with high entropy, no correlations, and the ability to protect symmetric, asymmetric, and hybrid encryption systems. In addition, real-life examples of chaotic models in cryptography, such as logistic mapping, the Lorentz system, and hybrid maps, were analysed to assess their effectiveness in generating random sequences, encrypting images, and expanding the key space. The critical analysis also identified the limitations of chaos-based generators and ways to improve their reliability, which is the basis for the practical use of chaotic systems in cryptography. The task of this stage was to determine the conditions under which chaos-based generators can be used in cryptography without losing their stability, as well as to formulate practical requirements that compensate for the lack of strict mathematical guarantees of their security.

In the fourth stage, the method of comparative analysis was used to compare chaotic generators and traditional PRNGs according to key criteria – nature, entropy, periodicity, predictability, speed, implementation features and application in cryptography, which determined their advantages and limitations. These criteria were chosen because they determine the suitability of a generator for cryptography. Nature reflects the principle of operation and the source of randomness, entropy and periodicity characterise the quality of the generated sequences, predictability is directly related to attack resistance, speed and implementation determine practical efficiency, and application in cryptography shows real applicability in data protection protocols. This identified the strengths and weaknesses of both approaches: to show the advantages of chaotic systems in providing high entropy, practical unpredictability and long periods, to highlight the disadvantages and to assess their suitability for use in cryptosystems. The methods of theoretical generalisation, critical analysis and comparative review of scientific sources were used to analyse and compare the models.

Results and Discussion

Nonlinear dynamic systems and their cryptographic applications in random number generation

A nonlinear dynamical system is defined as a mathematical model of processes in which the state change depends not only on time but also on previous values, and there are nonlinear relationships between the variables. Such systems are usually described by systems of differential equations or mappings, where the output is not proportional to the input. Characteristic properties include the presence of multiple equilibrium states, the ability to transition to unstable modes, self-organisation and the formation of complex behaviour even based on simple rules. Nonlinearity is the key source of chaotic regimes, in which the system demonstrates a complex and almost unpredictable evolution (Ming *et al.*, 2023).

In the field of cryptography, a fundamental aspect is the quality of random numbers used at all stages of data protection. In "symmetric encryption algorithms" (Advanced Encryption Standard (AES), ChaCha20, Data Encryption Standard (DES)), chaotic generators can be used to generate key material. The initial conditions and parameters of chaotic maps define a wide key space, which ensures the uniqueness and cryptographic strength of the obtained values. In addition, chaos can be used to generate the initialisation vectors required in Cipher Block Chaining (CBC), Counter Mode (CTR) and Galois/Counter Mode (GCM), where the randomness of the Initialisation Vector (IV) is a critical condition for protection against repetition. As proven by M.J.A. Calderon *et al*. (2024), in stream ciphers (ChaCha20-Poly1305), the uniqueness of the nonce is crucial to prevent reuse of the key stream, and it is the sensitivity of chaotic systems to initial conditions that can achieve such uniqueness.

In "asymmetric algorithms" (Rivest-Shamir-Adleman (RSA), Elliptic Curve Cryptography (ECC), as well as post-quantum schemes Kyber, Saber) with a public key, chaos can act as an auxiliary source of randomness. In particular, random values derived from chaotic processes can be used to generate seeds and parameters in RSA or ECC-based schemes (in Optimal Asymmetric Encryption Padding (OAEP)), which eliminates the problem of determinism. Additionally, chaotic maps are suitable for generating one-time random numbers required in key exchange protocols such as Diffie-Hellman and Elliptic Curve Diffie-Hellman (ECDH), increasing the resilience of systems to prediction (Garipcan *et al.*, 2025).

Chaotic permutations and maps are used in steganography to form complex patterns of data placement in images and multimedia files. This makes it much more difficult to detect hidden messages. Additionally, chaotic processes ensure the creation of watermarks that are resistant to attacks aimed to delete or modify. At the level of hardware implementations, chaotic oscillators, in particular the Chua circle, in combination with cryptographic extractors, can function as TRNGs (Nazish *et al.*, 2025). Such solutions are particularly relevant for embedded systems, sensor networks, and IoT devices, where the combination of high entropy and low power consumption is a critical requirement.

In practice, "hybrid cryptosystems" are mostly used, where asymmetry is used to securely transmit a symmetric key, and the data is encrypted directly using a symmetric

algorithm. The security of such systems directly depends on the entropy level of the initial random numbers from which the key material is formed. A critical component of cryptographic protocols is "random number generators". Insufficient entropy of keys (less than 128 bits) makes the system vulnerable to brute-force attacks. In the opinion M.J.A. Calderon *et al.* (2024), repeated nonces or IVs create conditions for key stream replay and recovery attacks. The lack of proper randomness in authentication protocols opens the way to replay attacks, and the predictability of the "k" parameter in digital signature schemes (Elliptic Curve Digital Signature Algorithm (ECDSA), Digital Signature Algorithm (DSA)) can lead to compromise of the private key.

Random number generators form the basis of modern cryptography of the 21st century, since the quality of their work directly determines the stability of cryptographic algorithms. The use of chaotic systems as a source of randomness is appropriate and fits seamlessly into the context of symmetric and asymmetric encryption, as well as additional applications, such as steganography and hardware solutions. In chaotic dynamic systems, attractors are central – sets to which trajectories tend over time. The so-called strange attractors, which have a fractal structure and multidimensionality, are fundamental; they determine the behavioural patterns of the system, not being reduced to regular periodicity. Another fundamental feature is the sensitivity to initial conditions: even minor changes in the initial parameters lead to fundamentally different development trajectories. This phenomenon, known as the "butterfly effect", directly contributes to the unpredictability of chaotic processes (Ding *et al.*, 2024). At the same time, a chaotic system remains deterministic, but due to the exponential growth of errors due to nonlinearity, its long-term behaviour cannot be accurately predicted.

Among the most common mathematical and physical models of chaos, there are several that are studied in the context of cryptography. The logistic mapping is a classical one-dimensional model of chaotic dynamics that describes the population dynamics of a system with limited resources. It can be used to generate pseudo-random numbers due to its ability to demonstrate chaotic behaviour at certain parameter values. The mathematical expression of the logistic mapping is as follows (May, 1976):

$$x_{n+1} = r x_n (1 - x_n), \quad (1)$$

where: x_n – value at the n-th step (the state of the system at time n); r – parameter that controls the dynamics of the system; x_{n+1} – value at the next step. This model can exhibit chaotic behaviour at certain values of the parameter , and its simplicity makes it common for applications such as pseudorandom number generation in cryptography.

A general discrete chaotic system is a system described by recurrent equations that reflect the relationship between the next and previous state of the system and has the form (Poincaré, 2017):

$$x_{n+1} = f(x_n, \theta), \quad (2)$$

where: θ – parameter that can be used to control a system, for example, it can affect the level of chaos or the transition between regular and chaotic behaviour; $f(x_n, \theta)$ – function that determines how a previous state of the system affects the next. Such a system is characterised by sensitivity to initial conditions, which is one of the main properties of chaos. Even a small change in the parameter θ can cause significant changes in the behaviour of the system.

The E.N. Lorenz (1963) system is a classic example of a three-dimensional nonlinear dynamical system consisting of three differential equations. It was developed for modelling atmospheric convection, but eventually became an icon of chaos theory due to its interesting and unpredictable properties, in particular, due to the E.N. Lorenz attractor, which shows how the system transitions between different states, and has the following form (3-5):

$$\dot{x} = \sigma (y - x); \quad (3)$$

$$\dot{y} = x (p - z) - y; \quad (4)$$

$$\dot{z} = xy - \beta z, \quad (5)$$

where: $\dot{x}$, $\dot{y}$, $\dot{z}$ – time derivatives that describe changes in each of the system variables; x – horizontal velocity (or convection); y – temperature or heat flux; σ – determines the rate of differentiation of the variable x relative to y, i.e. convection rate or flow rate; p – determines the temperature gradient or temperature difference between different layers of the atmosphere; β – describes vertical flows in atmospheric models or systems that demonstrate convection phenomena, z – vertical flow or vertical motion in systems. All parameters in the Lorentz system determine the interaction between temperature, convection rate and vertical flows in the medium. Changing any of these parameters can lead to a change in the behaviour of the system, from stable to chaotic. The Lorentz system is an example of a complex physical model that can be applied both at the mathematical level and in real-world cryptographic applications, such as key generation and encryption protocols.

As an electronic circuit, the Chua circle can be used to implement chaotic oscillations at the hardware level. It consists of an inductor, two capacitors, and a nonlinear resistive element (the so-called Chua diode), which can be used to exhibit a wide range of chaotic modes. The simplicity of the design and the possibility of physical implementation make this scheme a promising candidate for the creation of hardware chaotic signal generators that can be used in cryptography and information security systems (Alibraheemi *et al.*, 2024).

One of the key properties of chaotic systems is topological mixability. In phase space, this is manifested in the fact that any initial region of trajectories is eventually distributed over the entire domain of the definition. The system's trajectory visits an arbitrary neighbourhood with

a non-zero probability, which, for cryptography, means that there are no local patterns in the generated sequences, which ensures an even distribution of information and increases resistance to cryptanalysis. Another critical characteristic is high entropy and a rich state space. Entropy is a measure of disorder, which can be quantitatively described using formula (6) by C.E. Shannon (1948):

$$H(X) = -\sum_{ip}(x_i)\, log_2\, p\,(x_i), \quad (6)$$

where: $H(X)$ – entropy of a random variable X, which is a measure of disorder or the amount of information contained in a sequence of values X; $p\,(x_i)$ – probability of occurrence of the i-th element x_i in the sequence (the frequency of occurrence of a particular value in the data set); $log_2\, p\,(x_i)$ – logarithm of the probability of x_i in base 2, determines the number of bits required to encode the value x_i and expresses how much information this element contains. In chaotic systems, even short time series exhibit entropy values close to the maximum, making them similar to random sequences. This, in turn, provides a wide key space that is beyond the reach of brute force or effective cryptanalysis.

Another property is the long periods and the practical lack of repeatability. While traditional pseudo-random number generators have a finite period after which the sequence is reproduced, for multidimensional chaotic systems, this period can be so long that it is considered infinite from a practical point of view. This ensures the uniqueness of each generated sequence and makes it impossible to use repetition-based attacks. The fundamental basis for the use of chaos in cryptography is also provided by mathematical theorems confirming unpredictability, in particular, the positive value of the largest O.M. Lyapunov (1892) index:

$$\lambda = \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} ln|f'(x_k)|, \quad (7)$$

where: λ – largest Lyapunov exponent, a numerical criterion of the system's chaotic nature; $\lim_{n\to\infty}$ – shows that the assessment is conducted for a very large number of iterations, i.e. in the long run; $\frac{1}{n}\sum_{k=0}^{n-1}$ – average of all iterations k from 0 to n-1, averages local indicators of trajectory divergence; $ln|f'(x_k)|$ – natural logarithm of the modulus of the derivative of the function of the current state , measures the local divergence of neighbouring trajectories in phase space. The combination of a virtually infinite period and a positive Lyapunov exponent ensures that chaotic systems generate unique and unpredictable sequences, making them effective for cryptographic use.

The theorem of H. Poincaré's (2017) recurrence theorem proves that any trajectory of a system sooner or later returns to an arbitrarily small neighbourhood of the initial point, but the moment of return is fundamentally unpredictable (8, 9):

$$\forall U \supseteq \{x_0\}, \exists\{t_n\}: \varphi\,(t_n, x_0) \in U; \quad (8)$$

$$\forall n \in U, t_n \to \infty, \quad (9)$$

where: U – any neighbourhood of the initial point x_0; t_n – time after which the system trajectory enters the neighbourhood of U; $\varphi(t_n, x_0)$ – function that describes the position of the system at the moment of time t_n. Poincaré's theorem describes a critical property of chaotic systems: although their behaviour at any given time is unpredictable, they demonstrate recurrence, i.e. the ability to return to a certain region of phase space, albeit after a long time. This property is essential for cryptography, as it can be used for the generation of long, unpredictable and unique sequences to be used as keys.

The property of topological transitivity is one of the main characteristics of chaotic systems, which states that trajectories in the system eventually cover the entire region of phase space. In mathematical terms, this property is noted as (Gottschalk & Hedlund, 1955):

$$\varphi\,(t, U) \cap V \neq \emptyset. \quad (10)$$

This means that regardless of where a trajectory starts in phase space, it will eventually become in some other region of space. The system can thus "distribute" its behaviour throughout the phase space, which is a key characteristic of chaos. In the context of cryptography, this means that, based on any initial condition, the generated sequences will be evenly distributed throughout the entire space of possible values, which ensures unpredictability and a high level of entropy. This increases the resistance of cryptographic systems to attacks, as it becomes almost impossible to predict the long-term behaviour of a generator based on a short sequence.

An abstract model of key generation based on chaotic systems involves three stages: quantisation of chaotic values, cryptographic extraction, and updating the internal state. At the first stage, the continuous values obtained from the chaotic system are converted into a discrete form by scaling and rounding them to integer values. Formally, this process can be expressed as (Knuth, 1969):

$$u_n = \lfloor 2^w \times h(x_n) \rfloor, \omega = 64 \text{ or } 128, \quad (11)$$

where u_n – discretised value; $h(x_n)$ – function that determines the chaotic value at the n-th step; ω – number of bits for scaling accuracy; 2^w – scaling of values using a power of two. ω can be 64 or 128, which means the number of bits used to determine the accuracy and magnitude of the scaling. The value ω determines how many bits will be used to store each generated value. In the following stages, after obtaining the discretised values, they are cryptographically extracted using hash functions such as SHA-256 or others, which obtained key material or pseudorandom sequences. This process ensures the creation of a high-calorie, attack-resistant key by eliminating statistical dependencies and levelling the bit distribution, increasing resistance to cryptanalysis (Menezes *et al.*, 2011).

The discretised values u_n are used as input to the *SHAKE256* cryptographic hash function along with service

parameters (nonce and iteration index) and appear as (SHA-3 Standard, 2015):

$$y_n = SHAKE256(u_n \| nonce \| I), \quad (12)$$

where *SHAKE*256 – cryptographic hash function that generates random bits based on input values. It has the properties of high resistance to cryptanalysis and can generate arbitrarily long output bit sequences, y_n – result of the *SHAKE256* hash function, which are uniformly distributed random bits that can be used in subsequent stages of cryptographic algorithms. This stage ensures that the distribution is levelled, statistical dependencies are eliminated, and uniform random bits are obtained.

The internal state is updated based on the previous state and new values of the results to ensure the absence of repeated sequences and increase resistance to cryptographic attacks, in particular, prediction-based attacks. At this stage, the internal state of the system S_t is changed using a hash function that includes both the previous state S_t and the new random bits y_n obtained from the previous stage (using the *SHAKE256* hash function) (Barker & Kelsey, 2015):

$$S_{t+1} = \mathrm{H}(S_t \| y_n), \quad (13)$$

$$(\theta, x) \leftarrow MapFrom(S_{t+1}), \quad (14)$$

where: S_t – current internal state of the system at the t-th step; *MapFrom* – function that generates new values of parameters and state variables based on the new value of S_{t+1}. This process ensures that key sequences are unrepeatable, as updated parameters and state variables generate new trajectories in phase space each time. As a result, even if the cryptographic key is used repeatedly, its complexity and unpredictability remain at a high level. Thus, this stage guarantees high entropy and the absence of predictable patterns in the generated sequences, which makes the keys resistant to cryptanalysis and increases the system's reliability.

The results of the study showed that chaotic maps are capable of generating sequences with high entropy and statistical randomness. This result was consistent with the work of M. Irfan & M.A. Khan (2024), where they proposed a cryptographically secure generator based on a robust chaotic tent map. The study proved that their model demonstrates positive Lyapunov performance and thus meets the criterion of sensitivity to initial conditions. Testing according to the NIST SP 800-22 and TestU01 standards confirmed the statistical randomness of the output bit sequences. This confirmed the notion that chaotic systems can provide high entropy and cryptographic strength of generators.

The results of the study showed the effectiveness of chaos in steganography and multimedia data protection. This correlates with the study by Z.B. Madouri *et al.* (2024), developing a new pseudorandom generator based on chaotic digital filters. The study employed it to build an image encryption algorithm and proved that the generated sequences have high entropy and uniform distribution. The test results confirmed the algorithm's resistance to statistical attacks, which indicates the absence of noticeable correlations in chaotic trajectories. Additionally, the suitability of this approach for practical use in multimedia applications requiring reliability and speed was emphasised.

D. Murillo-Escobar *et al.* (2024) studied two chaos-based generators implemented on microcontrollers. The study tested their performance and the statistical randomness of the generated sequences. The test results showed compliance with NIST SP 800-22 criteria, which confirms the cryptographic suitability of the proposed solutions. A significant observation was that the implementation on resource-limited devices provides high entropy while reducing power consumption. This corresponded to the energy efficiency and practicality of chaotic generators in sensor networks and IoT systems outlined in the current study, confirming the feasibility of their use in lightweight cryptography.

Y. Alloun *et al.* (2025) presented a Field-Programmable Gate Array (FPGA) implementation of a generator that combines chaotic maps with artificial neural networks. The study emphasised that the integration of different approaches improves resistance to cryptanalysis. The experimental results confirmed the uniformity of the generated sequences and their compliance with NIST requirements, which indicates their cryptographic suitability. This correlated with the results of the study, which emphasised the need for hybrid cryptosystems and hardware implementations. This alignment demonstrated the feasibility of using chaos as an additional source of entropy in combination with other cryptographic primitives and confirmed the prospects of hardware solutions based on chaotic dynamic systems in cryptography.

The results of the study showed that long periods, recurrence and unpredictability are fundamental characteristics of chaotic systems in cryptography. In this context, the study by V. Patidar & T. Singh (2025) examined these properties in detail, proposing a new approach to random number generation based on Hamiltonian conservative chaotic systems. The study used Poincaré sections to generate non-periodic sequences that exhibit an almost infinite period. Verification using NIST SP 800-22 standards confirmed the cryptographic reliability of the proposed generator. This correlated with the current research on the butterfly effect, Poincaré's theorem, and the practical non-repeatability of chaotic trajectories, and serves as further evidence of the importance of the mathematical properties of chaos for creating cryptographically strong random number generators.

Thus, due to their fundamental properties, chaotic dynamical systems can be a reliable basis for generating cryptographic randomness. Their use in symmetric, asymmetric, and hybrid algorithms ensures high entropy, unpredictability, and cryptographic strength of key sequences. Practical research confirms the effectiveness of chaotic systems in cryptographic protocols and the prospects for further development of lightweight solutions.

Principles of construction and practical analysis of chaotic models of cryptographic key generation

For clarity, it is advisable to present the operation of the proposed model in the form of a flowchart showing the main stages of the process, from the generation of chaotic sequences to the formation of a cryptographic key. Such an approach can be used to trace the relationship between the individual stages of the model and determine the logic of its functioning. The flowchart also serves as a tool for further analysis and optimisation, as it demonstrates which system components are central in ensuring cryptographic security (Fig. 1).

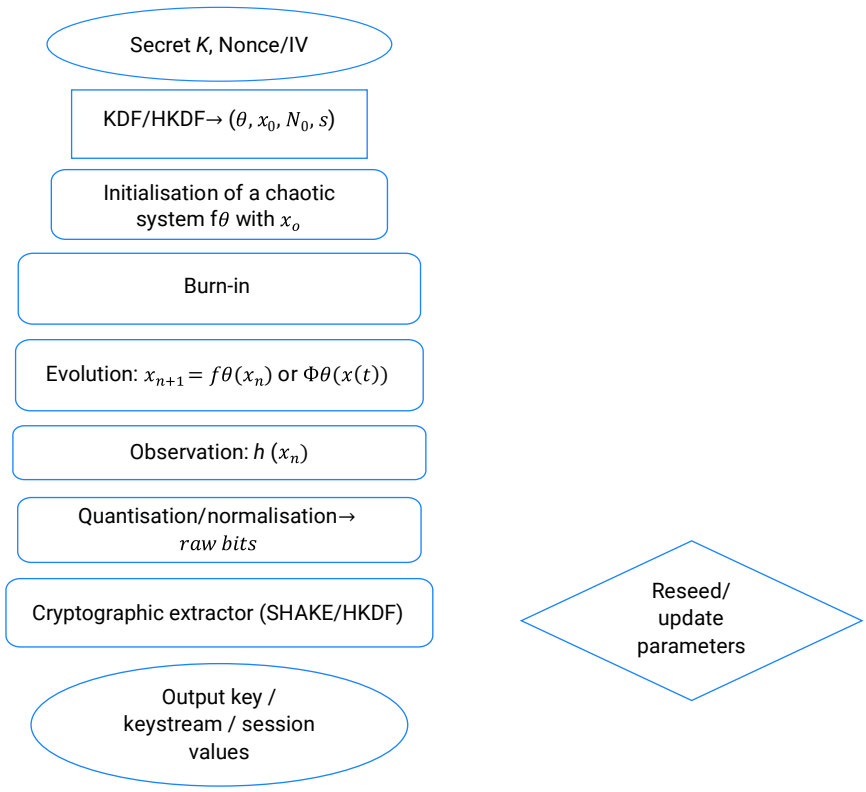


Figure 1. Flowchart of an abstract model of chaos-based key generation

Note: Secret K – secret parameters of the system that define the initial conditions of the chaotic model (Formulas 1-5); Nonce/IV – one-time random value, required for the uniqueness of the key stream (Formula 12); KDF/HKDF (θ, x_0, N_0, s) (Formulas 2, 11-14) – special function converts secret data into initial parameters of the chaotic system; Initialisation of the chaotic system f – starts the chaotic map or Lorentz system according to Formulas 1-5; Burn-in – discards initial iterations to eliminate transient effects (Formula 2); Evolution: $x_{n+1} = f\theta(x_n)$ or $\Phi\theta(x(t))$ – creation of a system trajectory based on recurrent equations (Formula 2); Obervation: $h(x_n)$ – mapping the internal state to the output data to be quantised (Formula 11); Quantisation/Normalisation (Formula 11) – continuous values into bit sequences; Cryptographic Extractor (SHAKE/HKDF, Formula 12) – levelling the distribution and removing statistical dependencies; Reseed / parameter update (Formulas 13,14) – update the internal state to ensure forward/backward secrecy; Output – key, keystream or session values for further use in cryptosystems (Formulas 12(result), 13,14)

Source: compiled by the author based on O.M. Lyapunov (1892), C.E. Shannon (1948), E.N. Lorenz (1963), D.E. Knuth (1969), A.J. Menezes *et al.* (2011), National Institute of Standards and Technology (2015), E. Barker & J. Kelsey (2015), H.M.M. Alibraheemi *et al.* (2024)

The principle of building a chaotic key model is to combine the dynamics of chaotic systems with cryptographic primitives. The initial conditions and parameters of the system are used as secret data. In the case of the logistic mapping, the key parameters are the coefficient r and the initial value x_0, while in the Lorenz system, the set of values (x_0, y_0, z_0) and the parameters σ, ρ, β. The size of the key should be sufficient to provide a space that is at least 2^{128}, as this is the minimum condition to prevent a complete search (Ming *et al.*, 2023). Next, the system evolves, resulting in a trajectory of successive states. To eliminate transient effects, burn-in is used, i.e. discarding the first iterations. The resulting values are subject to quantisation, i.e. conversion into bit sequences by scaling, mantissa extraction, or combinatorial operations (e.g. Exclusive OR (XOR)). The final step is cryptographic extraction (*SHAKE256*, HKDF), which removes patterns and ensures a uniform distribution of the output bits (Yin *et al.*, 2024).

The proposed scheme gives the model a number of properties. Firstly, the high entropy of chaotic sequences provides a level of disorder, which is quantified by the Shannon entropy. Secondly, the use of an extractor eliminates statistical correlations, making the original data appears as random data (Yin *et al.*, 2024). The sensitivity to initial conditions and parameters ensures that the same sequence cannot be reproduced without knowing the secret values, which is formally described by positive Lyapunov exponents. Updating the internal state ensures

forward and backward secrecy, and the use of hash functions with extended output makes the model scalable, capable of generating an arbitrary amount of key material with unchanged randomness characteristics (Tiwari *et al.*, 2025). Chaotic processes provide a full cycle of cryptographic key generation, from parameters and trajectories to bit sequences and their cryptographic amplification, creating a practically applicable basis for symmetric, asymmetric and hybrid encryption systems. An abstract model of chaos-based key generation combines the properties of chaotic systems with cryptographic extractors, providing a high level of entropy, uniform bit distribution, and no correlations. Updating the internal state increases resistance to compromise, while unpredictability and sensitivity to initial conditions make the model suitable for use in cryptosystems (Table 1).

Table 1. Examples of the application of chaotic generator models in cryptography

Model	Characteristic	Key objective	Results
Logistics mapping	An improved logistic map with "infinite chaos" for generating bit sequences is proposed. Implementing a logistics map as a PRNG in an FPGA	Develop a high-speed PRBG suitable for lightweight cryptography and IoT	The generator has passed NIST SP 800-22 tests; high entropy and no correlations were demonstrated. Demonstrated effectiveness for IoT and mobile devices
Lorenz system	The current cipher with a key stream generated by the Lorentz system is developed. A lightweight image encryption algorithm using Lorentz trajectories	Build a real cryptographic algorithm on a chaotic system. Ensure effective and fast image protection on mobile devices	The algorithm has shown high resistance to cryptanalysis and performance in a real-world environment. High entropy, uniformity of histograms, and resistance to statistical attacks have been achieved. Suitable for resource-limited systems
Hybrid models	Integration of Henon and Logistic maps to complicate chaotic trajectories	Improve image encryption security by expanding the key space	The algorithm has been proven to increase the key space and increase resistance to statistical and correlation attacks; it outperforms individual models

Note: PRBG – Pseudorandom Binary Generator
Source: compiled by the author based on M.J.A. Calderon *et al.* (2024), H.M.M. Alibraheemi *et al*. (2024), P.K. Singh *et al*. (2024), M. Nazish *et al*. (2025), A. Al-Hyari *et al*. (2025)

Chaotic systems have not only theoretical but also practical value in cryptography. Logistic mapping has been successfully used in both software and hardware implementations, providing high performance and entropy. The Lorentz system has been proven to be suitable for building both stream ciphers and lightweight algorithms for protecting multimedia data. The integration of several chaotic maps, as in the case of Henon and Logistic, demonstrates the possibility of further strengthening cryptographic strength by combining different models. Experimental results have confirmed the viability of chaotic generators as an alternative to classical DRBGs in various applications.

Nevertheless, it is necessary to strictly observe practical safety precautions in chaos-based generators. The use of raw chaotic trajectories without additional post-processing leads to correlations and statistical dependencies. To eliminate these problems, modern approaches use hash functions and "feedback key" mechanisms that ensure a uniform distribution of output bits and are confirmed by the results of NIST SP 800-22 tests (Yin *et al.*, 2024). In addition, the risk of "dead zones" in the parametric ranges of chaotic maps, which reduce entropy and narrow the key space, is emphasised. To solve this problem, it is proposed to use variable structures, for example, in the "structure-varying CML" model, which demonstrates a stable level of entropy and resistance to prediction (Ming *et al.*, 2023). The ultimate accuracy of calculations can cause hidden periodicity, so it is necessary to use high bit depth (64/128 bits) and regular state updates (reseeding) through cryptographic hashes. In addition, comprehensive testing is recognised as a mandatory stage of verification of chaos-based generators, which, in addition to the standard NIST SP 800-22/90B sets, includes analysis of autocorrelation functions, min-entropy estimation, and spectral methods (Ding *et al.*, 2024). Thus, scientific research has confirmed that the recommendations for the use of extractors, a wide key space, periodic state updates, and statistical testing are not only theoretically sound but also experimentally verified.

At the same time, in contrast to classical cryptographic primitives (RSA, ECC, AES-DRBG), where the security of algorithms is formally proved by reducing them to complex mathematical problems, chaos-based generators do not have security proofs. Their reliability is mostly confirmed by experimental tests (Ding *et al.*, 2024), entropy analysis (Nazish *et al.*, 2025), or numerical simulations (Calderon *et al.*, 2024), but not by formal mathematical reductions. This creates a certain gap between theory and practice, which is still under debate. Therefore, it is advisable to consider chaotic systems not as a full-fledged independent cryptographic mechanism, but as an additional source of entropy in hybrid schemes, where the final randomness is enhanced by cryptographic extractors (SHAKE, HKDF) and thus partially compensates for the lack of strict mathematical guarantees (Alibraheemi *et al.*, 2024; Singh *et al.*, 2024).

Both classical pseudorandom number generators and new approaches based on chaotic dynamical systems are used in cryptography. Classical PRNGs have advantages in terms of speed and ease of implementation, but their entropy and resistance to prediction are determined only by the quality of the algorithm. Instead, chaotic generators

use the properties of nonlinear systems, sensitivity to initial conditions, unpredictability, and almost infinite periods, which make them promising for generating cryptographic keys, nonces, and initialisation vectors (Table 2).

Table 2. Comparison of chaotic generators with traditional random number generators

Criteria	Chaotic generators	Traditional PRNG
Nature	Based on dynamic systems with nonlinear behaviour	Algorithmic designs, mainly linear or combined (linear congruent, Mersenne Twister, DRBG)
Entropy	High due to sensitivity to initial conditions and parameters, confirmed by Shannon's entropy	Depends on the algorithm; classic PRNGs have lower entropy
Frequency	Very long or almost infinite periods (in multidimensional systems)	Limited, the period depends on the bit depth
Predictability	Theoretically deterministic, but practically unpredictable due to chaos	Classic PRNGs are often predictable (if the state is known)
Performance	May be lower (mainly in differential models)	High performance, optimised for CPU/GPU
Implementation	Require precise numerical methods or hardware support (FPGA, analogue circuits)	Easy software execution
Application in cryptography	Promising for generating keys, nonces, and initialisation vectors	Standard DRBGs (AES-DRBG, Hash-DRBG, ChaCha20)

Source: compiled by the author based on H. Ming *et al.* (2023), M.J.A. Calderon *et al*. (2024), H.M.M. Alibraheemi *et al.* (2024), F. Yin *et al*. (2024), A. Tiwari *et al.* (2025), M. Nazish *et al.* (2025)

Chaotic generators demonstrate key advantages: high entropy and unpredictability due to their sensitivity to initial conditions, long periods and unique sequences, as well as the possibility of efficient software and hardware implementation. At the same time, their limitations are the loss of randomness due to finite accuracy and the lack of rigorous security proofs, which require the use of cryptographic extractors. The optimal approach is to integrate chaotic systems as an additional source of entropy in combination with classical cryptographic extractors. In practical applications, it is worth considering hardware implementations based on FPGAs and chaotic oscillators, in particular, Chua circles, which are especially relevant for IoT and sensor networks. To maintain cryptographic security, it is recommended to use regular reseeding, increased bit depth, and comprehensive statistical testing (NIST SP 800-22/90B, min-entropy analysis, spectral methods).

The results of the study showed that chaos-based generators require strict adherence to practical security considerations, as the use of raw chaotic trajectories without post-processing leads to correlations and statistical dependencies that need to be eliminated using hash functions, feedback key mechanisms, and comprehensive testing. This is consistent with the study by A. Sambas *et al.* (2024) on a dynamic analysis of a new three-dimensional chaotic system, which showed that only after careful optimisation of parameters and verification with statistical tests, PRNG demonstrates the required level of cryptographic reliability. The study emphasised the importance of spectral properties and recurrence to avoid hidden correlations. This means that the reliability of chaos-based generators can only be achieved by combining mathematical modelling, post-processing, and comprehensive testing.

Y. Alghamdi *et al.* (2022) proposed a lightweight image encryption algorithm based on a chaotic map and random substitution. The study emphasised that their approach is specifically designed for resource-constrained environments, such as mobile devices and IoT systems, where high performance and minimal power consumption are required. Experimental testing has shown that the generated sequences have high entropy, uniform histograms, and no statistical correlations. Additionally, it was proven that the algorithm demonstrates resistance to cryptanalytical attacks, including statistical and differential attacks. This correlated with the current study, which used logistic maps and the Lorenz system to build lightweight algorithms for encrypting multimedia data. This confirmed the practical feasibility of chaotic models in cryptography and proved their effectiveness in protecting information in real-world conditions.

The results of the study showed that chaotic oscillators are fundamentally different from traditional PRNGs, as they provide high entropy, long periods, and practical unpredictability, although they require complex numerical methods or hardware support for implementation. This correlates with the study by Y. Luo *et al.* (2024), who presented an FPGA implementation of a high-speed oscillator based on an n-dimensional chaotic system. The study proved that such a hardware implementation combine the characteristic properties of chaos with high performance and uniformity of output sequences. The tests have confirmed the cryptographic suitability of the generator, which proved the practical feasibility of using chaotic PRNGs in real security protocols.

The results of the study confirmed that the construction of a chaotic key generation model requires combining the dynamics of nonlinear systems with cryptographic primitives, including quantisation, extraction, and internal state updating, which ensures uniform bit distribution and forward/backward secrecy. This correlates with the study by M.A. Hadjadj *et al.* (2025), proposing a hardware implementation of PRNG-CS for embedded security systems. The authors emphasised the need to integrate chaotic dynamic processes with cryptographic post-processors to eliminate statistical correlations and increase cryptographic strength. Testing following NIST standards demonstrated

the high entropy and performance of the generator even in resource-limited environments and showed that chaotic models can be effectively implemented at the hardware level, ensuring reliable generation of key material.

M.T. Gençoğlu *et al.* (2025) presented a chaotic random number generator based on the quantum wave equation. The study emphasised that the use of chaotic dynamic properties alone is not sufficient to ensure cryptographic reliability. Therefore, they combined chaotic processes with post-processing mechanisms that eliminate correlations and guarantee a uniform distribution of the output bits. Testing has confirmed that this approach meets modern cryptographic requirements. This correlates with the results of the current study, where the integration of chaos with cryptographic extractors was recommended as the optimal solution. This proved the feasibility of combining nonlinear dynamics and classical cryptographic methods to create secure key generators.

Thus, chaos should be considered not as a self-sufficient crypto-primitive with formal security proofs, but as an additional entropy in standardised key generation circuits. This bridges the gap between chaos theory and cryptography practice and outlines a route to implementing chaos-based generators in real-world encryption, signature, and steganography protocols. In the future, this may contribute to the creation of more resilient and energy-efficient cryptographic systems.

Conclusions

The results of the study have shown that chaotic dynamical systems demonstrate a number of properties that make them promising in cryptography. The key characteristics include the presence of strange attractors with a fractal structure, sensitivity to initial conditions, and the butterfly effect, which cause unpredictability and high variability of the output sequences. Positive Lyapunov indices confirmed the exponential dependence of the trajectories on the initial parameters, which ensures that they cannot be reproduced exactly without knowing the secret values. Topological mixability and transitivity prove that chaotic trajectories are distributed uniformly in the phase space, which minimises local patterns and increases resistance to cryptanalysis. The presented model of key generation, which includes the stages of quantisation, hash extraction and internal state update, has confirmed the ability to eliminate statistical correlations and guarantee forward and backward secrecy.

Chaotic generators demonstrated high entropy values even on short time series, and their very long periods ensure the uniqueness of each sequence. The results confirmed the potential applicability of such generators in lightweight image encryption algorithms, which have shown resistance to cryptanalysis and efficiency in resource-constrained environments. The integration of chaotic maps, such as Henon and logistic maps, significantly expands the key space and increases the resistance to statistical attacks, making hybrid models a promising area of development. Chaotic generators can act as an additional source of entropy in modern cryptosystems, increasing their resilience and providing new directions for the development of information security theory and practice.

The limitation of the study was the theoretical nature and the lack of experimental verification of the results obtained, which requires further practical validation. Further research should focus on combining several chaotic maps, developing variable structures such as structure-varying CML, and expanding the range of tests, including the analysis of autocorrelation functions, min-entropy, and spectral characteristics.

Acknowledgements

None.

Funding

The study was not funded.

Conflict of Interest

None.

References

[1] Alghamdi, Y., Munir, A., & Ahmad, J. (2022). A lightweight image encryption algorithm based on chaotic map and random substitution. *Entropy*, 24(10), article number 1344. doi: 10.3390/e24101344.

[2] Al-Hyari, A., Abu-Faraj, M., Obimbo, C., & Alazab, M. (2025). Chaotic hénon-logistic map integration: A powerful approach for safeguarding digital images. *Journal of Cybersecurity and Privacy*, 5(1), article number 8. doi: 10.3390/jcp5010008.

[3] Ali, M., Ahmad, J., Khan, M.A.H., Ullah, S., Rehman, M.U., Shah, S.A., & Khan, M.S. (2025). A chaotic image encryption scheme using novel geometric block permutation and dynamic substitution. In F. Saeed, F. Mohammed, E. Mohhamed, S. Basura & M. Al-Sarem (Eds.), *Proceedings of the 4th international conference of advanced computing and informatics: Advances on intelligent computing and data science II* (pp. 1-12). Cham: Springer. doi: 10.1007/978-3-031-91351-8_1.

[4] Alibraheemi, H.M.M., Al Ibraheemi, M.M.A., & Radhy, Z.H. (2024). Design and practical implementation of a stream cipher algorithm based on a Lorenz system. *Journal of Information Security*, 4(3), 136-151. doi: 10.58496/MJCS/2024/019.

[5] Alloun, Y., Kifouche, A., Azzaz, M.S., Madani, M., Bourennane, E.-B., & Sadoudi, S. (2025). Design and FPGA implementation of a novel cryptographic secure pseudo random number generator based on artificial neural networks and chaotic systems. *Integration*, 103, article number 102388. doi: 10.1016/j.vlsi.2025.102388.

[6] Barker, E., & Kelsey, J. (2015). *Recommendation for random number generation using deterministic random bit generators*. Gaithersburg: U.S. Department of Commerce. doi: 10.6028/NIST.SP.800-90Ar1.

[7] Belazi, A., Kharbech, S., Aslam, N., Talha, M., Xiang, W., Iliyasu, A.M., & El-Latif, A.A.A. (2022). Improved Sine-Tangent chaotic map with application in medical images encryption. *Journal of Information Security and Applications*, 66, article number 103131. doi: 10.1016/j.jisa.2022.103131.
[8] Calderon, M.J.A., Lucas, L.J.L., Rosli, S.A.B., Ying, S.S.H., Lim, J.L.E., Xiang, M., & Teo, T.H. (2024). Logistic map pseudo random number generator in FPGA. *ArXiv*. doi: 10.48550/arXiv.2404.19246.
[9] Dhopavkar, T.A., Nayak, S.K., & Roy, S. (2022). IETD: A novel image encryption technique using tinkerbell map and duffing map for IoT applications. *Multimedia Tools and Applications*, 81, 43189-43228. doi: 10.1007/s11042-022-13162-x.
[10] Ding, P., Zhu, J., & Zhang, J. (2024). A four-dimensional no-equilibrium chaotic system with multi-scroll chaotic hidden attractors and its application in image encryption. *Physica Scripta*, 99, article number 105211. doi: 10.1088/1402-4896/ad7237.
[11] Faure, E., Shcherba, A., Skutskyi, A., & Lavdanskyi, A. (2024). A software model to generate permutation keys through a square matrix. *Bulletin of Cherkasy State Technological University*, 29(2), 10-23. doi: 10.62660/bcstu/2.2024.10.
[12] Fernández-Caramès, T.M., & Fraga-Lamas, P. (2020). Towards post-quantum blockchain: A review on blockchain cryptography resistant to quantum computing attacks. *IEEE Access,* 8, 21091-21116. doi: 10.1109/ACCESS.2020.2968985.
[13] Garipcan, A.M., Aydin, Y., & Özkaynak, F. (2025). An efficient 2D hyper chaos and DNA encoding-based s-box generation method using chaotic evolutionary improvement algorithm for nonlinearity. *Chaos, Solitons & Fractals*, 191, article number 115952. doi: 10.1016/j.chaos.2024.115952.
[14] Gençoğlu, M.T., Karaduman, Ö., & Özkaynak, F. (2025). Chaotic real number generator with quantum wave equation. *Symmetry*, 17(3), article number 349. doi: 10.3390/sym17030349.
[15] Gottschalk, W.H., & Hedlund, G.A. (1955). *Topological dynamics*. Providence: American Mathematical Society.
[16] Hadjadj, M.A., Kaibou, R., & Sadoudi, S. (2025). Design and hardware implementation of a prng-cs for embedded security applications. In *Proceedings of the 2025 IEEE computer society annual symposium on VLSI* (pp. 1-4). Los Alamitos: IEEE. doi: 10.1109/ISVLSI65124.2025.11130211.
[17] Irfan, M., & Khan, M.A. (2024). Cryptographically secure pseudo-random number generation (CS-PRNG) design using robust chaotic tent map (RCTM). *ArXiv*. doi: 10.48550/arXiv.2408.05580.
[18] Jackson, J., & Perumal, R. (2025). A robust image encryption technique based on an improved fractional order chaotic map. *Nonlinear Dynamics*, 113, 7277-7296. doi: 10.1007/s11071-024-10480-7.
[19] Knuth, D.E. (1969). *The art of computer programming*. Reading: Addison-Wesley.
[20] Krulikovskyi, O., Haliuk, S., Ivashko, V., & Politanskyi, R. (2024). Periodicity of timeseries generated by logistic map: Part II. *Security of Infocommunication Systems and Internet of Things*, 2(2), article number 02003. doi: 10.31861/sisiot2024.2.02003.
[21] Kumar, A., & Dua, M. (2021). Novel pseudo random key & cosine transformed chaotic maps based satellite image encryption. *Multimedia Tools and Applications*, 80, 27785-27805. doi: 10.1007/s11042-021-10970-5.
[22] Kumar, M., & Ch, D. (2025). Enhancing image security through a fusion of chaotic map and multi-level scrambling techniques. *Signal, Image and Video Processing*, 19, article number 235. doi: 10.1007/s11760-025-03814-4.
[23] Lorenz, E.N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2), 130-141.
[24] Luo, Y., Fan, C., Xu, C., & Li, X. (2024). Design and FPGA implementation of a high-speed prng based on an n-D non-degenerate chaotic system. *Chaos, Solitons & Fractals*, 183, article number 114951. doi: 10.1016/j.chaos.2024.114951.
[25] Lyapunov, O.M. (1892). *General problem of stability of motion*. Kharkiv: Zilberberga's typography.
[26] Madouri, Z.B., Said, N.H., & Pacha, A.A. (2024). A new pseudorandom number generator based on chaos in digital filters for image encryption. *Journal of Optics*, 53, 3548-3563. doi: 10.1007/s12596-023-01606-y.
[27] May, R.M. (1976). Simple mathematical models with very complicated dynamics. *Nature*, 261, 459-467. doi: 10.1038/261459a0.
[28] Menezes, A.J., van Oorschot, P.C., & Vanstone, S.A. (2011). *Handbook of applied cryptography*. Boca Raton: CRC Press.
[29] Ming, H., Hu, H., & Zheng, J. (2023). Design and application of a structure-varying coupled chaotic system with high security. *Expert Systems with Applications*, 226, article number 120158. doi: 10.1016/j.eswa.2023.120158.
[30] Murillo-Escobar, D., Vega-Pérez, K., Murillo-Escobar, M.A., Arellano-Delgado, A., & López-Gutiérrez, R.M. (2024). Comparison of two new chaos-based pseudorandom number generators implemented in microcontroller. *Integration*, 96, article number 102130. doi: 10.1016/j.vlsi.2023.102130.
[31] Nazish, M., Javid, M., & Banday, M.T. (2025). Enhanced logistic map with infinite chaos and its applicability in lightweight and high-speed pseudo-random bit generation. *Cybersecurity*, 8, article number 24. doi: 10.1186/s42400-024-00319-4.
[32] Patidar, V., & Singh, T. (2025). A novel approach to pseudorandom number generation using hamiltonian conservative chaotic systems. *Frontiers in Physics*, 13, article number 1553389. doi: 10.3389/fphy.2025.1553389.
[33] Poincaré, H. (2017). *The three-body problem and the equations of dynamics: Poincaré's foundational work on dynamical systems theory.* Cham: Springer. doi: 10.1007/978-3-319-52899-1.

[34] Sambas, A., Benkouider, K., Kaçar, S., Ceylan, N., Vaidyanathan, S., Sulaiman, I.M., Mohamed, M.A., Ayob, A.F.M., & Muni, S.S. (2024). Dynamic analysis and circuit design of a new 3D highly chaotic system and its application to pseudo random number generator (PRNG) and image encryption. *SN Computer Science*, 5, article number 420. doi: 10.1007/s42979-024-02766-9.
[35] SHA-3 Standard: Permutation-based hash and extendable-output functions. (2015). doi: 10.6028/NIST.FIPS.202.
[36] Shandyba, A. (2025). *Method of embedding digital watermarks using chaotic maps*. Kharkiv: Kharkiv National University of Radioelectronics.
[37] Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379-423. doi: 10.1002/j.1538-7305.1948.tb01338.x.
[38] Singh, P.K., Jha, B., & Kumar, S. (2024). An efficient and lightweight image encryption technique using Lorenz chaotic system. *Mathematical Modeling and Computing*, 11(3), 702-709. doi: 10.23939/mmc2024.03.702.
[39] Tiwari, A., Diwan, P., Diwan, T.D., Miroslav, M., & Samal, S.P. (2025). A compressed image encryption algorithm leveraging optimized 3D chaotic maps for secure image communication. *Scientific Reports*, 15, article number 14151. doi: 10.1038/s41598-025-95995-8.
[40] Yin, F., Li, A., Lv, C., Wu, R., & Gao, S. (2024). A new image encryption algorithm with feedback key mechanism using two-dimensional dual discrete quadratic chaotic map. *Nonlinear Dynamics*, 112, 20417-20435. doi: 10.1007/s11071-024-10099-8.

Застосування теорії хаосу для підвищення стійкості систем шифрування в інформаційних технологіях

Володимир Луханін

Кандидат фізико-математичних наук, асистент
Харківський національний університет радіоелектроніки
61166, просп. Науки, 14, м. Харків, Україна
https://orcid.org/0000-0003-4328-929X

Анотація. Метою дослідження було теоретичне обґрунтування застосування хаотичних динамічних систем для підсилення стійкості криптографічних ключів. Методологія дослідження базувалася на теоретичному, порівняльному та критичному аналізі наукових джерел для оцінки потенціалу хаотичних систем. Встановлено, що хаотичні карти забезпечують високу ентропію, довгі періоди та непередбачуваність згенерованих послідовностей завдяки чутливості до початкових умов, що підтверджується розрахунками ентропії Шеннона та позитивними Ляпуновськими показниками. Використання хеш-функцій та механізмів оновлення внутрішнього стану усуває статистичні кореляції й підвищує стійкість генераторів до криптоаналізу. Показано, що послідовності, отримані на основі логістичного відображення та системи Лоренца, проходять стандартні статистичні тести NIST SP 800-22, демонструючи рівномірність розподілу та відсутність кореляцій. Використання кола Чуа як аналогової схеми забезпечує фізично реалізовані генератори істинної випадковості (True Random Number Generator) з низьким енергоспоживанням, придатні для ресурсно-обмежених Internet of Things-систем. Схема з інтеграцією кількох хаотичних карт підтвердила збільшення простору ключів і підвищення стійкості до статистичних атак, у порівнянні з традиційними PseudoRandom Number Generator. Виявлено, що хаотичні генератори здатні забезпечити forward і backward secrecy завдяки оновленню внутрішнього стану системи, що запобігає повторюваності послідовностей. Хаотичні генератори мають переваги над традиційними завдяки дуже довгим періодам і чутливості до початкових умов, проте їх ефективність залежить від криптографічної постобробки та правильного вибору параметрів. Рекомендовано застосування хаотичних систем як додаткового джерела ентропії в програмних і апаратних реалізаціях, зокрема у легковагових криптографічних рішеннях для інтернету речей, сенсорних мереж і мобільних пристроїв. Практична значимість полягає у застосуванні результатів розробниками для безпечного шифрування, дослідниками для генерації випадкових чисел та інженерами інтернету речей для захисту пристроїв

Ключові слова: нелінійна динаміка; генератори випадкових чисел; криптографічна ентропія; хаотичні атрактори; ініціалізаційні вектори; топологічна транзитивність; криптографічна екстракція

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 10.08.2025 Revised: 22.10.2025 Accepted: 23.12.2025

UDC 004.65 DOI: 10.31649/vitce/3.2025.101

Comparison of data consistency models in distributed database management systems

Andrii Myrhorodskyi*

Postgraduate Student
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0009-0007-6764-0060

Oksana Romaniuk

PhD in Technical Science, Associate Professor
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0000-0003-0235-8615

Abstract. The use of distributed infrastructure to ensure scalability and high availability creates new challenges for maintaining data consistency between rapidly growing information system nodes that require reliable data management for correct operation. The aim of the study was to comprehensively systematise and comparatively analyse methods for ensuring data consistency in distributed database management systems, taking into account the fundamental trade-offs between consistency, availability and update delays described by the CAP and PACELC theorems. To achieve this goal, methods of theoretical analysis, formal modelling of system behaviour, and comparative expert evaluation were used. As a result of the study, consistency models were systematised according to two main approaches: data-centric and client-centric. The first approach analyses models that determine the global behaviour of the system: linearity, sequential, causal and eventual consistency. The advantages, disadvantages and typical application scenarios are identified for each model. The second approach considers client-oriented models that provide guarantees within a single user session: read and write consistency, monotonic read, monotonic write, and session causality. A generalised classification is proposed that visualises the relationship between the degree of consistency, delays, flexibility, fault tolerance and potential performance for each model. All considered data consistency models are compared using a number of selected essential characteristics (PACELC class, consistency, fault tolerance, potential performance, etc.) and diagrams based on their parameters. The practical value of the work lies in the formulation of clear recommendations for selecting the optimal consistency model depending on the requirements for reliability, performance, and architectural features of the information system. The results can be used to improve the efficiency of designing distributed databases in high-load systems, such as financial services, Internet of Things platforms, and cloud applications

Keywords: distributed systems; Data-Centric consistency; Client-Centric consistency; CAP theorem; PACELC; system availability; fault tolerance

Introduction

In modern conditions of rapid development in distributed computing systems and increasing demands for their speed and efficiency, the issue of ensuring data consistency in Distributed Database Management Systems (DDBMS) is becoming particularly relevant. Most contemporary information services utilise distributed infrastructure to achieve more flexible scalability, high availability, and improved fault tolerance. However, the use of these advantages is accompanied by a number of additional problems and limitations related to data consistency among the system's nodes, particularly during asynchronous replication, network delays, or the inoperability of individual components.

Scientific research in recent years has significantly expanded the theoretical and practical understanding of

Suggested Citation:
Myrhorodskyi, A., & Romaniuk, O. (2025). Comparison of data consistency models in distributed database management systems. *Information Technologies and Computer Engineering*, 22(3), 101-112. doi: 10.31649/vitce/3.2025.101

*Corresponding author

consistency models. For example, the work by P.S. Almeida (2024) proposed an axiomatic model for describing consistency in asynchronous distributed systems, which unites and generalises existing approaches. It also formulates the CLAM theorem for classifying system behaviour based on selected consistency guarantees: Closed past (finalised operations without the possibility of re-execution), Local visibility, Arbitration, and Monotonic visibility. According to the research, maintaining all four criteria for the complete set of data abstractions in systems without additional waiting is impossible. This offers an alternative perspective to the CAP (Consistency, Availability, Partition tolerance) and PACELC (Partition tolerance, Availability, Consistency, Else, Latency, Consistency) theorems, which can be used to derive and compare different consistency models.

A detailed overview of challenges regarding consistency and data integrity in DDBMS is presented in the work by H. Mahmoud & H. Yasin (2025), which focused on classic consistency protocols (such as 2PC) and modern replication methods. The Database Integrity (DDBI) project proposed by the authors offers an innovative solution through active triggers, rule-oriented integrity enforcement, and multi-version object tracking, which, according to the results obtained, is a better alternative to traditional approaches. Additionally, D. Russel *et al*. (2025) proposed an innovative approach to consistency verification in the context of serverless and edge architectures, which is extremely relevant in the transition to hybrid cloud environments. The authors used a comprehensive framework combining formal specifications, static analysis, real-time monitoring, and adaptive assurance, which also opens up prospects for using machine learning to predict consistency violations. Separate attention is warranted for the research by S. Ghasemirad *et al*. (2025), which implemented the Eiger-PORT+ protocol to ensure Transactional Causal Consistency with convergence (TCCv), confirmed by formal verification with TCCv isolation guarantees in Isabelle/HOL. The results refuted the previous hypothesis about incompatibility with transactional writes in the presence of performance-optimised read-only transactions. This represented the first full formal verification of a complex distributed database protocol and opened up the possibility of developing new abstract and practical models and protocols.

Concurrently, the article by Y. Chen *et al*. (2024) considered specific implementations of systems supporting various consistency models in cloud NoSQL platforms – from strong to bounded staleness. The work focused on supporting a wide range of load balancing, latency reduction, and resource management mechanisms popular in cloud environments for multi-model systems. In the work by N. Faria & J. Pereira (2025), the use of Conflict-Resistant SQL Views (CRDV) is proposed to ensure consistency in hybrid transactional and analytical workloads. An important distinction of this work is its focus on supporting CRDV in existing popular relational SQL databases, which allows for improved performance, greater flexibility in data merging strategies, and expands their use as heterogeneous DDBMS in large-scale information systems. Researchers also payed special attention to consistency algorithms in replicated systems, particularly CRDT (Conflict-free Replicated Data Types) approaches, which are reviewed in the works of Yu. Rabeshko & Yu. Turbal (2023), confirming the relevance of the topic in modern scientific discourse.

Thus, despite noticeable achievements in the field of consistency models, a need remains for a unified classification of existing approaches, a systematisation of their application conditions, and a determination of the algorithmic and engineering constraints imposed when implementing the respective models in DDBMS. Understanding these features can provide a better foundation for the further development of this area and the design of new data consistency models. The aim of this research was the formalisation and systematisation of methods for ensuring data consistency in DDBMS, taking into account the requirements of the CAP and PACELC theorems, as well as modern approaches to the implementation of Data-Centric and Client-Centric consistency models. Achieving this goal involved the execution of the following tasks: analysis and classification of modern consistency models; identification of the limitations and advantages of each approach in view of the system type; and determination of practical recommendations for model selection depending on usage scenarios.

Materials and Methods

The research is grounded in a theoretical and analytical method, which was used to study and compare data consistency models, as well as to analyse and interpret them within the context of the CAP and PACELC theorems. The initial stage of the study involved the collection and analysis of information about consistency models from relevant literature and available technical (or research) documentation. The models selected for the research were those that are widespread or have had a significant impact on the development of the DDBMS field, such as linearisability, which significantly influenced subsequent data-centric models. This stage also included the analysis of existing classification methods, such as the division into Data-Centric and Client-Centric models (proposed in the work by H.N.S. Aldin *et al*. (2019)), and methods for formally modelling the behaviour of distributed systems under various network conditions. This analysis was carried out based on the categorisation of systems according to the criteria of consistency (C), availability (A), partition tolerance (P), and update latency (L). For systematisation of the results, system types were categorised: CP, AP, CA (according to CAP), and PA/EL, PA/EC, PC/EL, PC/EC (according to PACELC (Abadi, 2012)).

Based on the obtained data, an analytical review of the consistency models was conducted, highlighting their core architectural principles, advantages, disadvantages, and typical application scenarios in distributed systems. This stage involved theoretical modelling of the Data-Centric and Client-Centric consistency models. A comparative method was used to contrast the characteristics and properties of the models, along with expert evaluation of their

parameters, including the level of consistency, flexibility, latency, fault tolerance, and potential performance. System-structural analysis was applied to build an independent classification of characteristics such as those primarily dependent on theoretical principles and those primarily dependent on implementation and to define the interrelationships between consistency models.

The next stage of the research was performed by visualising the acquired data in MS Excel to display the classification results and facilitate further analysis. To visualise the relationships between the models' properties, a generalised classification table of characteristics was used, which compares the model type, its correspondence to the PACELC theorem, and the evaluated expert parameters. The table allowed for the conceptualisation of the trade-offs between performance and consistency across different classes of DDBMS models, as well as trends in more modern models, such as the prioritisation of availability and latency according to the PACELC theorem. The use of radar charts (or spider diagrams) allowed for a graphical comparison of the characteristics' assessments obtained through theoretical analysis and enabled the tracking of their change and dependence on typical usage scenarios of modern information systems. The results of the analysis served as the basis for substantiating the choice of an appropriate consistency model type depending on the system's target application scenario, which allows for the adaptation of the DDBMS architecture to the specified requirements for availability, performance, or data integrity. The final stage of the research involved using the comparative method and content analysis to correlate the obtained data and methods for classifying characteristics with existing academic works in the field of DDBMS and distributed systems.

Results and Discussion

An information system consisting of only a single node is responsible for processing all possible read and write operations, whereas a distributed system has a defined set of nodes that do this concurrently. If a single-node system becomes inoperable, all operations become impossible to execute. One of the properties of distributed systems is the ability to continue functioning even when one or more nodes are inoperable. On the other hand, this creates the potential for node desynchronisation: network problems can interrupt the data replication process, and simultaneous write operations on different nodes may conflict with each other.

Data consistency is the property of a distributed system to maintain all nodes in a uniform and predictable state. Data consistency is closely related to and balances against other system properties and characteristics, such as performance and availability - ensuring consistency requires additional logic and can limit processing speed, but it provides specific guarantees regarding data synchronisation between nodes (Ahmed *et al.*, 2023). Data consistency mechanisms help order the execution of concurrent requests and provide options for system recovery from failures without data corruption.

The CAP and PACELC theorems are frequently used to describe the fundamental limitations in data consistency models in distributed systems. The CAP theorem (or Brewer's theorem) defines three key characteristics of distributed systems (Muñoz-Escoí *et al.*, 2019):

- Consistency: following any write operation, the distributed system must return the latest version of the data upon a read request, regardless of which node the request was made to (Lourenço *et al.*, 2015).
- Availability: the distributed system must provide a response to any incoming request, even if a portion of the nodes are inoperable at the moment the request is being processed.
- Partition tolerance: in the event of a loss of connection (or other network problems) between several nodes (splitting the cluster into several parts with limited communication), the distributed system must continue to operate and process incoming requests.

The CAP theorem states that a distributed system can only guarantee the simultaneous assurance of two out of the three described characteristics/guarantees:

- CP (Consistency & Partition Tolerance) systems will maintain data consistency in the event of network problems by limiting availability, i.e., the ability to process incoming requests. Without limiting availability, nodes with disrupted communication would be unable to synchronise completed write operations, which would lead to server desynchronisation and a violation of consistency.
- AP (Availability & Partition Tolerance) systems operate according to an algorithm opposite to CP systems nodes continue to process requests, ensuring availability even in the case of network problems, but the results of these requests may be inconsistent between nodes. AP systems often either limit the number of nodes accepting write requests or switch entirely to processing only read requests.
- CA (Consistency & Availability) systems always ensure consistency and availability and, therefore, cannot continue to operate if network problems occur either all nodes process requests and synchronise with each other, or continued operation is impossible. DDBMSs that support ACID transactions are typically included in this category of distributed systems.

The CAP theorem focuses its attention on the behaviour of a distributed system during cluster partitioning (when communication problems between nodes occur). Under normal network conditions, a system can maintain both consistency and availability simultaneously, regardless of its conditional type (CP, AP, or CA). Changes in behaviour and state occur only at the moment network problems arise. The PACELC theorem extends the CAP theorem by adding a new classification of behaviour in the absence of system partitioning due to communication problems. If there are network problems in the distributed system and the system can potentially be partitioned or partially inoperable, the choice described in the CAP theorem (between data consistency and availability) is applied. However, if

the system is operating under normal network conditions, the choice is applied between consistency (C) and update latency (L) (Golab, 2018). These relationships are visually represented in the Figure 1.

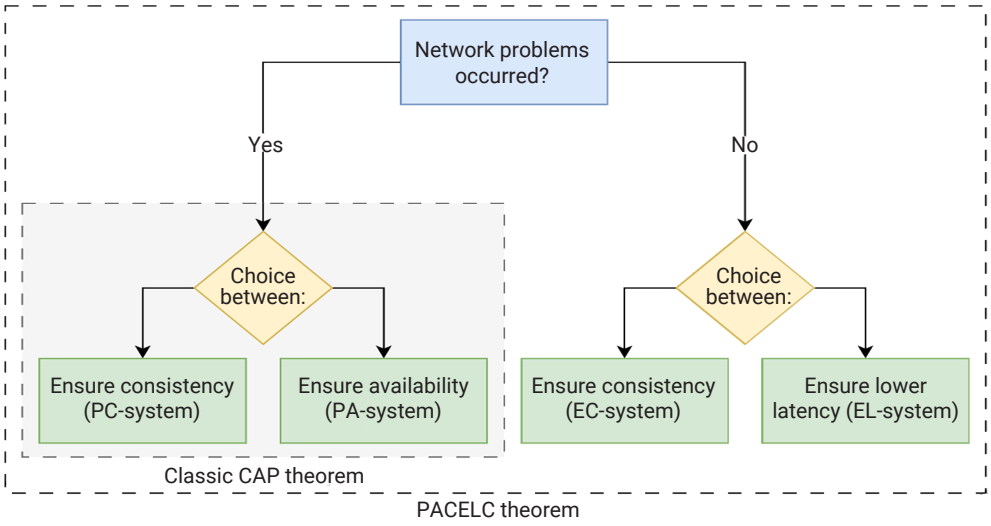


Figure 1. Choice in the PACELC theorem
Source: developed by the authors based on research by D. Abadi (2012)

If a system prioritises consistency, any data replication must include a guarantee that all nodes in the distributed system are updated before request processing continues. That is, a data write operation is considered fully complete when all nodes have received/replicated that operation. This guarantees the strict consistency required for ACID transactions but can decrease the speed (increase latency) of query execution due to the need for communication among all system nodes.

Conversely, if the distributed system prioritises lower latency and greater speed, an operation is considered complete after being written to just one or a few nodes (depending on the system's implementation). Node synchronisation in this scenario is considered a separate, asynchronous operation that can occur independently of subsequent write operations. Accordingly, some nodes in such a distributed system may be inconsistent with the rest of the cluster at certain times and return stale data.

Combining the behaviour options with and without network problems, the PACELC theorem divides distributed systems into four types (Gorbenko *et al.*, 2020):

- PA/EL: the system prioritises Availability during communication problems, and Lower Latency during normal functioning.
- PA/EC: the system prioritises Availability during network partitioning, but otherwise prioritises Consistency over lower latency.
- PC/EL: the system prioritises Consistency during network partitioning, but favours Lower Latency under normal conditions.
- PC/EC: the system prioritises Consistency both during network problems and under normal functioning.

It should be noted that the classification of both theorems is not rigid – some systems support several types of behaviour depending on the choice of the user or administrator of the DDBMS. For example, Amazon DynamoDB can be classified as a PA/EL system, i.e., availability and lower latency are prioritised over consistency, but the user may require strict data consistency when executing queries – in such cases, DynamoDB will function as a PA/EC system. In addition to classification according to CAP and PACELC theorems, there is also a division into data-centric and client-centric data consistency models (Aldin, 2019). Data-centric consistency models are built around the rules for replicating data in a distributed system as a whole. Such models describe the required behaviour of the entire system and its states. For example, an data-centric model may include rules for replicating operations, their ordering, and the logic for simultaneously processing read or write requests.

Client-centric consistency models describe the rules of operation within a single client or session, rather than the system as a whole. Such models typically provide consistency guarantees for operations from a single user, but the global state of the distributed system or the same operations for another user may be temporarily inconsistent. For example, a distributed system that prioritises lower latency can guarantee the consistency of all operations for a single client – if the same client changes the data and then reads it immediately, it will always receive the latest version. However, its operations will not be immediately replicated to other nodes.

Here are some examples of well-known methods for ensuring data consistency and their classification. Linearizability is an data-centric model that guarantees strict data consistency. This model is based on the global ordering of all operations of a distributed system into a single list. The actual process of system operation may involve several clients working in parallel with the same data, but

the model assumes that between receiving a request and responding to it, there is a moment when the operation is considered to be performed instantaneously or atomically (Park *et al*., 2024). Also, the time of sending the response to the request is used to order events, rather than the time of starting its processing. Using these two concepts, it is possible to transform any history of requests of a real distributed system into a single ordered list of events, which acts as a source of truth for all nodes and thus ensures strict data consistency. Implementing linearizability in an DDBMS will require the use of a data locking mechanism or another system for synchronous replication of operations between nodes, such as a consensus algorithm. Accordingly, such a system will belong to the PC/EC category according to the PACELC theorem.

Sequential consistency is another data-centric model that is similar to linearisability and orders operations into a single sequence. The main difference lies in the absence of ordering based on the completion time of the request execution. The global history must include operations from all nodes in the order they were executed on the respective server, but the ordering of operations from different nodes can be arbitrary and does not necessarily reflect their true execution order (Perrin *et al*., 2016). Thus, the sequential model provides less strict data consistency compared to linearisability. This simplification potentially offers higher performance for the distributed system, which is well-suited for distributed NoSQL databases used as a distributed cache. Accordingly, distributed systems based on the sequential model can be classified as PC/EL models under PACELC.

Causal consistency is a data-centric consistency model that only guarantees the preservation of the sequence of interrelated operations. For example, write and read requests for two different, unrelated table rows do not affect each other. Therefore, a database management system based on causal consistency may not maintain a strict order of their execution this will not affect the final state of the database or change the result of the user's operation (Junfeng *et al*., 2022). However, if the system has information about a potential link between operations, such as reading and updating the same value, their sequence must be preserved and correctly replicated to all system nodes to maintain consistency. In such a case, this model will function similarly to linearisability or sequential consistency.

This flexible approach gives causal consistency even greater potential operational performance, making it well-suited for DDBMSs with less strict consistency requirements. This model is more challenging to classify under PACELC, as its flexibility allows for more freedom in software implementation. Generally, a system based on causal consistency is less vulnerable to network problems and does not require forced node synchronisation unless the operations are related, so it can be classified as PA/EL.

Another well-known data-centric model is eventual consistency. It offers the weakest data consistency guarantees of all the models discussed so far, as it does not inherently require operations to be stored in a specific sequence or ordered on individual system nodes (Xu *et al*., 2024). Eventual consistency establishes a synchronisation rule for the data: in the absence of new updates, the distributed system will eventually reach the same state on all nodes, meaning it will become consistent. During request processing and until the synchronisation of all nodes is complete, a distributed database using eventual consistency may return stale data. Typically, such systems also have additional mechanisms for writing/updating data to avoid potential conflicts between nodes during concurrent operations.

Thus, eventual consistency only guarantees that the distributed system will be consistent at some point in time but does not impose stricter conditions or limitations on when this will occur. This simplification provides the highest potential request processing speed among all the data-centric models reviewed. Furthermore, eventual consistency is well-suited for information systems with high availability requirements, which is why it is used in NoSQL DDBMSs, such as Amazon DynamoDB. Accordingly, distributed systems based on eventual consistency can also be classified into the PA/EL class according to the PACELC theorem. A generalised comparison of the properties and application of the reviewed data-centric models is provided in Table 1.

Table 1. Comparison of data-centric consistency models

No.	Model name	Description	Advantages	Disadvantages	Application
1	Linearisability	Ensures a strict global ordering of all operations based on the time of the request response	The strictest guarantees of data consistency	High latency, complex implementation, low performance	Banking systems, critical transaction management systems, and systems with mission-critical data
2	Sequential consistency	Ensures a strict ordering of requests within a single server, but without coordinating the order between nodes	Simpler implementation than linearisability; strict data consistency	Does not guarantee the real-time order of operations in a distributed cluster	NoSQL databases, caching systems with low consistency requirements
3	Causal consistency	Guarantees the ordering of only causally related operations. Independent operations can be executed in any order	Simpler implementation; no need for global synchronisation	The need to track causal links; complexity of the logic depends on the data structure	Collaborative applications, social networks, and distributed collaborative editing environments
4	Eventual consistency	The system eventually reaches a uniform state (synchronises) across all nodes, but new updates may not be immediately reflected	Highest performance, minimal latency	Lack of strict guarantees; a possibility of receiving outdated data	Web applications, mobile applications, and caching systems.

Source: developed by the authors

Client-centric consistency models typically describe the rules for the interaction of a client (or several clients) with a single node in the system. Thus, consistency guarantees are provided only for requests sent specifically to that node, while the distributed system as a whole may be in an inconsistent state or synchronise according to other rules. The read-your-writes model guarantees that if a request to write/update data has been successfully executed, all subsequent requests to read data will return the updated value corresponding to the previous changes. The main difference between this model and others is the need to maintain the order of operations for only one client on a given node; there is no global ordering at the time of the request, and the system may appear inconsistent to other clients until synchronisation occurs (Aldin *et al.,* 2020). Due to the absence of global synchronisation rules, the read-write consistency model is somewhat similar to the eventual consistency model, but provides more stable behaviour within the operations of a single client (as long as that client is connected to the same node). Therefore, this model is well suited for DDBMSs for data caching or for information systems where a large amount of information is tied to specific users and the risks of conflicts during global synchronisation are lower. According to PACELC, this model can be classified as PA/EL.

The monotonic reads model is a client-centric model that guarantees data reads within a single session. If the client is connected to the same node, all data read operations will be consistent – the results of new queries cannot return an older version of information than that already provided in previous queries (Campêlo *et al.,* 2020). Similar to the read-write consistency model, such a distributed system does not require global ordering of operations and does not restrict the data synchronisation process – it can be asynchronous and initiated by an individual node as needed. On the other hand, the monotonic read model does not provide any guarantees for data update requests – they may appear inconsistent even within a single server. This model is well suited for implementing DDBMS for working with cache or information systems where most of the load consists of read operations. According to PACELC, this model can also be classified as PA/EL.

The monotonic write model is a data consistency model that guarantees the ordering of write operations according to the order in which the client initiated them (Taghinezhad-Niar, 2024). That is, if a client has made several writes in sequence, the order in which they will be written to the distributed system node will be exactly the same. It is important to consider some limitations of this model: first, ordering guarantees are only provided for operations of a single client, as in other client-centric models. That is, there is no global ordering, and the overall list of operations may differ on each node, but the operations of each individual node will have the correct order of execution. Second, this model does not provide consistency guarantees when reading data, so each client may receive both new and stale data in response to its query.

The monotonic record model, similar to other models of this class, can be classified as PA/EL type according to the PACELC theorem. A DDBMS with monotonic record consistency is most useful for applications with multi-user data editing, such as online document editors, so that individual users' edits are correctly ordered. However, this model can also be used in certain types of caching systems.

The Session Causality model (also known as Write Follows Reads Consistency) is a client-centric model that resembles a simplified version of the data-centric causal consistency model. While causal consistency orders all causally related operations at the level of the system as a whole, the Session Causality model tracks only the read-write operations of a single client (Viotti & Vukolić, 2016). The model makes the logical assumption that a data read operation might prompt the client to certain actions, including a data write operation. Accordingly, the write request must be processed on those distributed system nodes where the data is no older than what the client last read. This preserves the cause-and-effect relationship of operations within a single session. The session causality model can similarly be classified as PA/EL, and its application scenarios in distributed systems and DDBMSs are analogous to other client-centric models.

While data-centric models describe the logic of data consistency at the level of the entire distributed system, client-centric models give only limited guarantees concerning specific events, and in most usage scenarios, they do not conflict with each other. As a result, a distributed system may implement several different consistency guarantees at once this approach is distinguished as a separate model called Session consistency (Wang *et al*., 2024) or the Session consistency model. Typically, the session consistency model incorporates the guarantees of all four reviewed client-centric models: Read-your-writes, Monotonic Reads, Monotonic Writes, and Session Causality however, the actual implementation of the information system may include other or modified guarantees. Furthermore, it should be noted that all reviewed models only offer rules for data consistency within a session; the actual mechanisms for synchronising this data and resolving potential conflicts remain unaddressed. A generalised comparison of the properties and application of the reviewed client-centric models is provided in Table 2.

Table 2. Comparison of client-centric consistency models

No.	Model name	Description	Advantages	Disadvantages	Application
1	Read-your-writes	Guarantees local consistency within a user's session - after a write, a client only sees updated data	Guarantee of sequentiality for the user; easy to understand	Potential for conflicts between different clients	Systems with personal data, user profiles, and caching systems

Table 2. Continued

No.	Model name	Description	Advantages	Disadvantages	Application
2	Monotonic reads	Guarantees that a client will never read data older than what they were previously read	Stable consistency in reading	Does not guarantee correct writes; it is possible to get outdated information for other users	Analytical systems, reading logs, and statistical overviews
3	Monotonic writes	A single client's writes occur in the same order in which they were initiated.	Preserves the logic of changes.	No guarantees on reads; possible conflicts between clients	Online editors, version control systems
4	Session causality	Guarantees causal links between operations for a single client.	Strict consistency within a single session or user	Complex to implement, latency during relationship checks	Collaborative editing of documents, cloud platforms
5	Session consistency	A comprehensive model that combines all previous client-centric models	Highest local consistency; preserves client interaction logic	Does not guarantee global consistency; dependent on the user's session	Interactive web applications, personalised services, and mobile applications that maintain session state

Source: developed by the authors

The comparisons in Tables 2 and 3 were conducted based on the models' main advantages, disadvantages, and typical application scenarios. Such an analysis allows for the identification of development trends, particularly when comparing consistency models within the same class. For instance, in both tables, later models show not only improvements in functional characteristics, such as increased speed or local consistency, but also an orientation towards use in mobile and web applications. Since most models have varying degrees of formalisation without a golden standard implementation, it's impossible to compare them using specific numerical values. However, the existing classification systems for these consistency models can be expanded by relatively comparing individual characteristics and properties.

The model type comparison uses the classical division into data-centric and client-centric models. This characteristic should be considered first when comparing the models against each other, as, although most client-centric models offer a high degree of consistency, this consistency is limited to the session between the system node and the client. Only a data-centric model can fully ensure data consistency purely between the nodes of a distributed system. The PACELC Class describes the models' behaviour according to the PACELC theorem, allowing them to be quickly classified based on their actions under normal operation and in the presence of network problems. Other characteristics use relative assessments based on available theoretical data and application examples. The defined characteristics and metrics of the models are presented in Table 3.

Table 3. Characteristics of data consistency models

Consistency model	Model type	PACELC class	Degree of consistency	Latency	Flexibility	Fault tolerance	Potential performance
Linearisability	Data-centric	PC/EC	High	High	Low	Medium	Low
Sequential consistency	Data-centric	PC/EL	High/ Medium	Medium	Medium	Medium	Medium
Causal consistency	Data-centric	PA/EL	Medium	Low	High	Medium	Medium
Eventual consistency	Data-centric	PA/EL	Low	Low	High	High	High
Read-your-writes	Client-centric	PA/EL	Locally medium	Low	Low	Medium	High
Monotonic reads	Client-centric	PA/EL	Locally medium	Low	Low	Medium	High
Monotonic writes	Client-centric	PA/EL	Locally medium	Low	Low	Medium	High
Session causality	Client-centric	PA/EL	Locally high	Medium	Low	Medium	High
Session consistency	Client-centric	PA/EL	Locally high	Low	Medium	Medium	High

Source: developed by the authors

According to the assessments, the data-centric models demonstrate a gradation from strict consistency guarantees with high latency to weak guarantees with improved performance, which can also be linked to the changing application scope of the respective DDBMSs in Tables 1-2. client-centric models achieve a locally high level of consistency while maintaining low global latency, making them suitable for use in information systems without high data integrity requirements. It is worth noting that the majority of the reviewed models are classified as PA/EL, which indicates a greater focus on availability and low latency in modern DDBMSs. Furthermore, the selected metrics for relative comparison can be tentatively divided into two categories – those primarily dependent on theoretical principles and those primarily dependent on implementation, as shown in Figure 2.

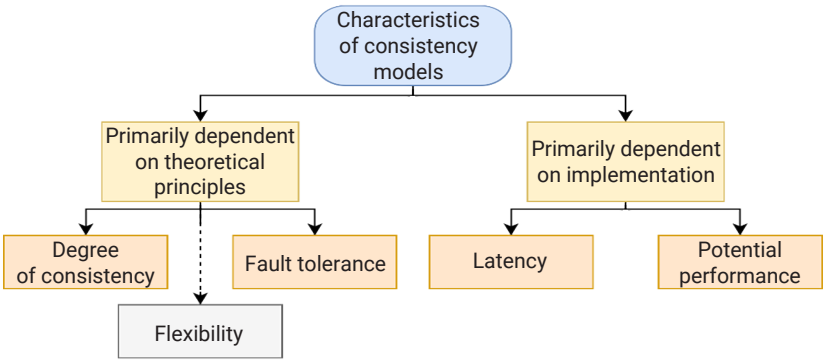


Figure 2. Classification of characteristics for comparing consistency models
Source: developed by the authors

This classification reflects only a tendency towards one side or the other. True (non-relative) assessments must be made only when a reference implementation is available for each of the data consistency models under review. For example, flexibility was classified as a characteristic primarily dependent on theoretical principles, but of all other metrics in this category, it is the most susceptible to the specifics of practical software implementation. Accordingly, it has a special designation on the diagram.

Degree of consistency and latency characterise the main parameters of the reviewed models what data consistency guarantees are provided between nodes and how strongly these guarantees affect performance. It is important to note that the assigned scores are relative and based on available theoretical knowledge, so they should only be used to compare the reviewed models with each other. Also, since client-centric models do not account for consistency with other nodes, a different measure of consistency is used for their classification.

Flexibility and fault tolerance characterise the potential ability of a distributed system with a certain type of consistency to adapt to changes (including the inoperability of individual nodes) and various usage scenarios. Automation of the recovery process, data replication frequency, and actual downtime are important factors when choosing both a disaster recovery strategy and the consistency model in which it will be used (Myrhodskyy *et al.*, 2023). Potential performance summarises all preceding characteristics for a relative comparison of the operational efficiency of different models against each other.

If scores from 1 to 3 are used instead of the "low"–"high" gradation for the reviewed characteristics, a radar chart can be constructed for each consistency model. This type of chart allows for a more visual comparison of the models a larger area on the radar chart signifies greater universality and a higher overall score relative to other consistency models. It should be noted that latency is a negative characteristic, so its correspondence between the verbal gradation and the scores is inverse – a model with high latency receives a score of 1. Due to the specifics of the degree of consistency in client-centric models, their comparison should be conducted separately. The radar charts for data-centric models are shown in Figure 3.

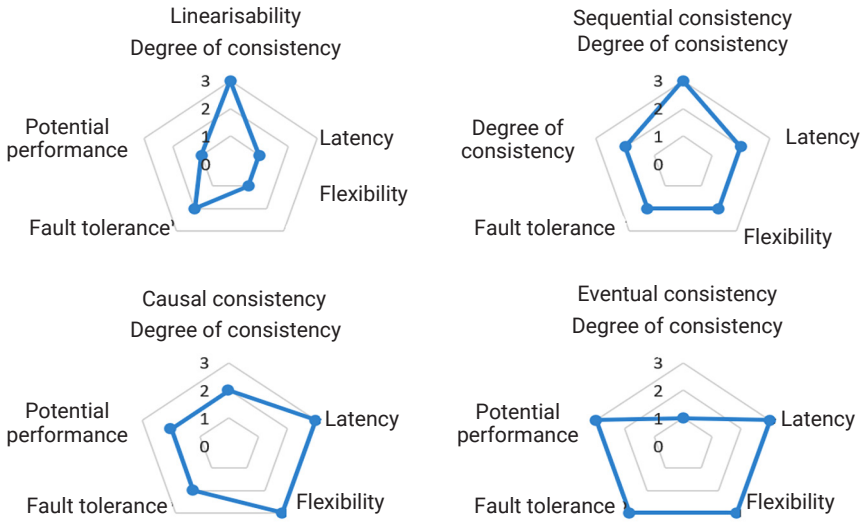


Figure 3. Comparison of data-centric consistency models
Source: developed by the authors

The charts allow to track changes in characteristics when transitioning to progressively weaker data-centric models. Linearisability, sequential consistency, and causal consistency are similar to one another, but each successive model reduces the complexity of the consistency guarantees. This, in turn, reduces implementation complexity and operation execution time, allowing for improvements in the models' latency, potential performance, and flexibility. Under these conditions, eventual consistency, with the weakest guarantees, shows the best potential performance and other characteristics provided that such a degree of consistency satisfies the requirements of the specific distributed information system. The radar charts for client-centric models, shown in Figure 4, display different relationships between characteristics.

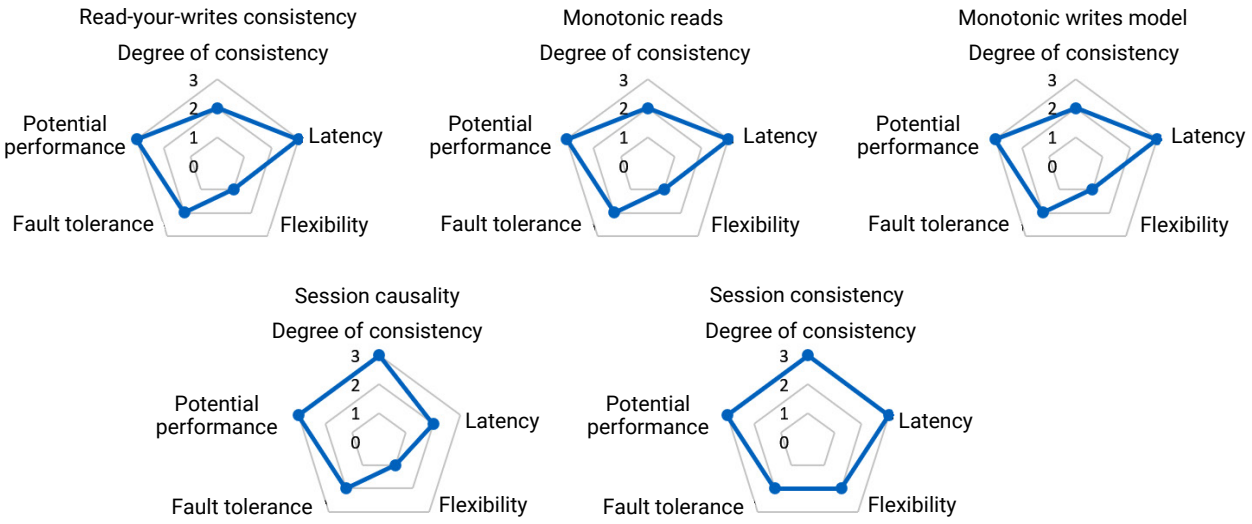


Figure 4. Comparison of data-centric consistency models

Source: developed by the authors

Read-your-writes, monotonic reads, and monotonic writes all exhibit similar metrics due to comparable constraints inherent in their architectural principles. Although they specialise in different application scenarios, their potential performance and flexibility remain similar under the best conditions for each model. The final two models, session causality and session consistency, stand out for having better characteristics due to their hybrid architecture. session causality has a higher degree of consistency due to its kinship with data-centric models, while the superior metrics of session consistency are possible only through the absence of additional constraints in client-centric models, allowing multiple approaches to be applied simultaneously to cover the widest range of application scenarios.

It is worth noting that while the comparison is theoretical, it allows to track a certain trend: older data consistency models with stricter guarantees are less flexible and less performant, whereas the best speed, even at a theoretical level, is offered by models and methods with the minimum degree of global consistency. Thus, maintaining data in a consistent state is a complex task that is directly interconnected with other properties and characteristics of the created distributed system. The researched information and theoretical foundations of existing consistency models can be used to develop a proprietary method for ensuring data consistency that can cover a broader spectrum of application scenarios.

The results obtained are particularly interesting in the context of comparison with the research by E. Brewer (2012), where the author of the original CAP theorem reviews its application in a modern context. The author's analysis determined that the balance between consistency and availability can vary depending on the DDBMS subsystem or the current operation. This suggests that classifying models according to intermediate consistency levels is appropriate, which aligns with the results obtained from the analysis of data-centric models, where the level of guarantees is reduced to speed up operations. However, the author emphasised the need to develop and integrate mechanisms for operability and recovery under network problems, which was only partially addressed in this work during the analysis of client-centric models.

A similar analysis of consistency models was performed by H.N.S. Aldin *et al*. (2019), which considers not only data-centric and client-centric models but also newer hybrid solutions. The authors examined only the general operating principles of the models and their architectural solutions, without the detailed identification and comparison of their characteristics performed in this study. However, their focus on additional DDBMS characteristics is important, such as issues of scalability, security, cost optimisation (which is particularly crucial in the context of using cloud providers), handling stale data, and the energy efficiency of modern distributed systems.

Similar results were obtained in the study by R. Cattell (2011), where the main focus was on comparing the SQL and NoSQL systems that use the consistency models, rather than the models themselves. The author conducted a broad analysis of key-value and document-oriented

DDBMSs, as well as considering the problems of classic relational systems. Based on this, the author predicted the continued proliferation of systems using a lesser degree of consistency to improve other characteristics - a similar trend was found during this research, exemplified by the development of data-centric systems. The analysis by Z. Mahfoud & N. Nouali-Taboudjemat (2019) also considers the classification of consistency models in the context of the CAP and PACELC theorems, but places a greater focus on existing distributed systems from the largest cloud providers, including Amazon Web Services, Google Cloud, and Microsoft Azure. Furthermore, the researchers analyse the possibilities of using multiple consistency models in a single DDBMS depending on the concept, which is represented in some software products, but a more detailed comparison of properties and characteristics, as performed in the current research, is absent.

The work by M. Diogo *et al.* (2019) is based on most of the reviewed consistency models, but with less focus on the data-centric sphere. Moreover, it analyses the practical implementation of these models in existing software products, including hybrid solutions like Cassandra, MongoDB, and Neo4j. This allowed the authors to gain a broader representation of consistency model application scenarios, but the resulting comparison was based solely on the CAP theorem, without a more detailed investigation of behaviour according to PACELC or other systems. An analysis and classification system similar to those used in this research can be found in the work by B. Pradeep (2023). The author investigated the features of the CAP theorem, its drawbacks, and similarly arrived at a classification using PACELC and a relative comparison of consistency models using the characteristics of latency, core operating principles, and application scenarios. A significant difference is the consideration of the coordination complexity of data-centric models when scaling DDBMSs. However, the author did not explore the application and comparison of client-centric consistency models.

The trade-off between the level of data consistency and system speed and availability identified in this research is also tracked in the work by D. Nguyen *et al*. (2019). The authors detail the operating principles and limitations of eventual and sequential consistency models, modelling various implementations of DDBMSs and their specific mechanisms in the context of performance improvement, which provides extensive practical data for further research. However, the analysis and comparison of other models with a weaker degree of consistency, as well as the possibilities of applying client-centric models which are particularly important in the context of key-value DDBMSs on which the work is focused remains unresolved.

Conclusions

In this article, a comprehensive analysis of methods for ensuring data consistency in distributed database management systems was carried out. Consistency serves as a key factor in ensuring the correct operation of such systems in a multi-node environment where both network failures and conflicting write operations are possible. It has been established that consistency mechanisms play a critical role in maintaining the reliability and predictability of data, especially with concurrent access and asynchronous replication.

As a result of the research, the main approaches to classifying consistency models were systematised based on the CAP and PACELC theorems, with an emphasis on practical trade-offs between consistency, availability, latency, and resistance to cluster partitioning. Data-Centric models (linearisability, sequential consistency, causal consistency, eventual consistency) and Client-Centric models (read-your-writes, monotonic reads, monotonic writes, session causality, session consistency) were analysed. Based on the analysis, their advantages, disadvantages, and application scenarios were determined, and a comparison was performed across a range of selected characteristics, such as the degree of consistency, fault tolerance, flexibility, latency, and potential performance. The comparison allowed to establish a relationship between the degree of consistency and performance among the data-centric models: the principles of linearisability and sequential consistency provide better consistency guarantees but suffer from higher latency and lower potential performance. At the same time, client-centric models have a narrower scope of application as they only provide data consistency guarantees within a session or within a single node and a few clients, but in such usage scenarios, they show good performance and performance.

The causal and eventual consistency models can be highlighted as the most promising for further research and development. The causal model shows a good balance between the selected characteristics while maintaining a sufficiently high degree of data consistency, and the eventual model offers potentially the best performance among data-centric models. Among the client-centric models, session consistency is worth noting, as it combines most of the characteristics of the other models in this category and is the most universal and effective. Due to a different synchronisation mechanism, this model is not suitable for resolving conflicts between nodes and cannot be directly compared with data-centric models, but its principles can be used to develop specific methods within dynamic or adaptive data consistency models. Further research should focus on finding ways to improve the performance and fault tolerance of the causal consistency model, enhancing the degree of consistency in the eventual model while preserving other metrics, and researching and developing adaptive and dynamic consistency models.

Acknowledgements

None.

Funding

The study was not funded.

Conflict of Interest

None.

References

[1] Abadi, D. (2012). Consistency tradeoffs in modern distributed database system design: CAP is only part of the story. *Computer*, 45(2), 37-42. doi: 10.1109/mc.2012.33.

[2] Ahmed, J., Karpenko, A., Tarasyuk, O., Gorbenko, A., & Sheikh-Akbari, A. (2023). Consistency issue and related trade-offs in distributed replicated systems and databases: A review. *Radioelectronic and Computer Systems*, 2(106), 171-179. doi: 10.32620/reks.2023.2.14.

[3] Aldin, H.N.S., Deldari, H., Moattar, M.H., & Ghods, M.R. (2019). Consistency models in distributed systems: A survey on definitions, disciplines, challenges and applications. *ArXiv*. doi: 10.48550/arXiv.1902.03305.

[4] Aldin, H.N.S., Deldari, H., Moattar, M.H., & Ghods, M.R. (2020). Strict timed causal consistency as a hybrid consistency model in the cloud environment. *Future Generation Computer Systems*, 105(C), 259-274. doi: 10.1016/j.future.2019.11.038.

[5] Almeida, P.S. (2024). A framework for consistency models in distributed systems. *ArXiv*. doi: 10.48550/arXiv.2411.16355.

[6] Brewer, E. (2012). CAP twelve years later: How the "rules" have changed. *Computer*, 45(2), 23-29. doi: 10.1109/mc.2012.37.

[7] Campêlo, R.A., Casanova, M.A., Guedes, D.O., & Laender, A.H.F. (2020). A brief survey on replica consistency in cloud environments. *Journal of Internet Services and Applications*, 11(1), article number 1. doi: 10.1186/s13174-020-0122-y.

[8] Cattell, R. (2011). Scalable SQL and NoSQL data stores. *ACM SIGMOD Record*, 39(4), 12-27. doi: 10.1145/1978915.1978919.

[9] Chen, Y., Pan, A., Lei, H., Ye, A., Han, S., Tang, Y., Lu, W., Chai, Y., Zhang, F., & Du, X. (2024). TDSQL: Tencent distributed database system. *Proceedings of the VLDB Endowment*, 17(12), 3869-3882. doi: 10.14778/3685800.3685812.

[10] Diogo, M., Cabral, B., & Bernardino, J. (2019). Consistency models of NoSQL databases. *Future Internet*, 11(2), article number 43. doi: 10.3390/fi11020043.

[11] Faria, N., & Pereira, J. (2025). CRDV: Conflict-free replicated data views. *Proceedings of the ACM on Management of Data*, 3(1), 1-27. doi: 10.1145/3709675.

[12] Ghasemirad, S., Sprenger, C., Liu, S., Multazzu, L., & Basin, D. (2025). Pushing the limit: Verified performance-optimal causally-consistent database transactions. In A. Gurfinkel & M. Heule (Eds.), *Tools and algorithms for the construction and analysis of systems. TACAS 2025. Lecture notes in computer science* (Vol. 15698, pp. 43-62). Cham: Springer. doi: 10.1007/978-3-031-90660-2_3.

[13] Golab, W. (2018). Proving PACELC. *SIGACT News*, 49(1), 73-81. doi: 10.1145/3197406.3197420.

[14] Gorbenko, A., Karpenko, A., & Tarasyuk, O. (2020). Analysis of trade-offs in fault-tolerant distributed computing and replicated databases. In *2020 IEEE 11th international conference on dependable systems, services and technologies (DESSERT)* (pp. 1-6). Kyiv: IEEE. doi: 10.1109/DESSERT50317.2020.9125078.

[15] Junfeng, T., Wenqing, B., & Haoyi, J. (2022). PGCE: A distributed storage causal consistency model based on partial geo-replication and cloud-edge collaboration architecture. *Computer Networks*, 212(C), article number 109065. doi: 10.1016/j.comnet.2022.109065.

[16] Lourenço, J.R., Cabral, B., Carreiro, P., Vieira, M., & Bernardino, J. (2015). Choosing the right NoSQL database for the job: A quality attribute evaluation. *Journal of Big Data*, 2(1), article number 18. doi: 10.1186/s40537-015-0025-0.

[17] Mahfoud, Z., & Nouali-Taboudjemat, N. (2019). Consistency in cloud-based database systems. *Informatica*, 43(3), 313-319 . doi: 10.31449/inf.v43i3.2650.

[18] Mahmoud, H.A., & Yasin, H.M. (2025). Data integrity and consistency challenges in distributed database systems. *Engineering and Technology Journal*, 10(5), 5077-5086. doi: 10.47191/etj/v10i05.36.

[19] Muñoz-Escoí, F.D., de Juan-Marín, R., García-Escrivá, J.-R., González de Mendívil, J.R., & Bernabéu-Aubán, J.M. (2019). CAP theorem: Revision of its related consistency models. *The Computer Journal*, 62(6), 943-960. doi: 10.1093/comjnl/bxy142.

[20] Myrhodskyy, A.V., Romanyuk, O.V., Romanyuk, O.N., & Titova, N.V. (2023). Development of a high availability method for configuration management software. *Optoelectronic Information-Power Technologies*, 46(2), 64-75. doi: 10.31649/1681-7893-2023-46-2-64-75.

[21] Nguyen, D., Charapko, A., Kulkarni, S.S., & Demirbas, M. (2019). Using weaker consistency models with monitoring and recovery for improving performance of key-value stores. *Journal of the Brazilian Computer Society*, 25(1), article number 10. doi: 10.1186/s13173-019-0091-9.

[22] Park, S., Kim, J., Mulder, I., Jung, J., Lee, J., Krebbers, R., & Kang, J. (2024). A proof recipe for linearizability in relaxed memory separation logic. *Proceedings of the ACM on Programming Languages*, 8(PLDI), 175-198. doi: 10.1145/3656384.

[23] Perrin, M., Petrolia, M., Mostéfaoui, A., Jard, C. (2016). On composition and implementation of sequential consistency. In C. Gavoille & D. Ilcinkas (Eds.), *Distributed computing. DISC 2016. Lecture notes in computer science* (Vol. 9888, pp 284-297). Berlin: Springer. doi: 10.1007/978-3-662-53426-7_21.

[24] Pradeep, B. (2023). Data consistency models in distributed systems: CAP theorem revisited. *International Journal on Science and Technology*, 14(3). doi: 10.5281/zenodo.14631471.

[25] Rabeshko, Yu. , & Turbal , Yu. (2023). Review of joint text editing algorithms Conflict-free Replicated Data Types (CRDT). *Bulletin of Cherkasy State Technological University*, 28(4), 10-18. doi: 10.62660/2306-4412.4.2023.10-18.

[26] Russel, D., Dawson, R., Chen, N., & Chambers, A. (2025). *Consistency models and verification in modern distributed systems*. Retrieved from https://www.researchgate.net/publication/389598133_Consistency_Models_and_Verification_in_Modern_Distributed_Systems.
[27] Taghinezhad-Niar, A. (2024). A client-centric consistency model for distributed data stores using colored petri nets. In *10th international conference on web research (ICWR)* (pp. 309-314). Tehran: IEEE. doi: 10.1109/ICWR61162.2024.10533365.
[28] Viotti, P., & Vukolić, M. (2016). Consistency in non-transactional distributed storage systems. *ACM Computing Surveys (CSUR)*, 49(1), article number 19. doi: 10.1145/2926965.
[29] Wang, C., Mohror, K., & Snir, M. (2024). Formal definitions and performance comparison of consistency models for parallel file systems. *IEEE Transactions on Parallel and Distributed Systems*, 35, 1092-1106. doi: 10.1109/TPDS.2024.3391058.
[30] Xu, Q., Yang, C., & Zhou, A. (2024). Native distributed databases: Problems, challenges and opportunities. *Proceedings of the VLDB Endowment*, 17(12), 4217-4220. doi: 10.14778/3685800.3685839.

Порівняння моделей забезпечення узгодженості даних у розподілених системах керування базами даних

Андрій Миргородський

Аспірант
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0007-6764-0060

Оксана Романюк

Кандидат технічних наук, доцент
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0003-0235-8615

Анотація. Використання розподіленої інфраструктури для забезпечення масштабованості та високої доступності створює нові виклики для підтримки узгодженості даних між вузлами інформаційних систем, що стрімко зростають та вимагають надійного управління даними для коректної роботи. Метою дослідження була комплексна систематизація та порівняльний аналіз методів забезпечення узгодженості даних у розподілених системах керування базами даних, враховуючи фундаментальні компроміси між узгодженістю, доступністю та затримками оновлення, що описуються теоремами CAP та PACELC. Для досягнення мети було використано методи теоретичного аналізу, формального моделювання поведінки систем та порівняльного експертного оцінювання. В результаті дослідження було систематизовано моделі узгодженості за двома основними підходами: інформаційно-орієнтованим та клієнтоорієнтованим. В рамках першого підходу проаналізовано моделі, що визначають глобальну поведінку системи: лінеаризовність, послідовну, причинну та кінцеву узгодженість. Для кожної моделі визначено переваги, недоліки та типові сценарії застосування. У рамках другого підходу розглянуто клієнтоорієнтовані моделі, що надають гарантії в межах сесії одного користувача: узгодженість читання і запису, монотонне зчитування, монотонний запис та сесійну причинність. Запропоновано узагальнену класифікацію, яка візуалізує співвідношення між ступенем узгодженості, затримками, гнучкістю, стійкістю до збоїв та потенційною продуктивністю для кожної моделі. Проведено порівняння усіх розглянутих моделей узгодженості даних за допомогою ряду відібраних істотних характеристик (клас за PACELC, узгодженість, стійкість до збоїв, потенційна продуктивність тощо) та діаграм на основі їх параметрів. Практична цінність роботи полягає у формулюванні чітких рекомендацій щодо вибору оптимальної моделі узгодженості залежно від вимог до надійності, продуктивності та архітектурних особливостей інформаційної системи. Результати можуть бути використані для підвищення ефективності проектування розподілених баз даних у високонавантажених системах, таких як фінансові сервіси, платформи Інтернету речей та хмарні застосунки

Ключові слова: розподілені системи; інформаційно-орієнтована узгодженість; клієнтоорієнтована узгодженість; CAP-теорема; PACELC; системна доступність; відмовостійкість

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 27.08.2025 Revised: 12.11.2025 Accepted: 23.12.2025

UDC 620.179.1 DOI: 10.31649/vitce/3.2025.113

Methods of signal processing and data interpretation for detecting microdefects in industrial materials

Kanan Mikayilov*

Postgraduate Student
Azerbaijan State Oil and Industry University
AZ1010, 20 Azadliq Ave., Baku, Azerbaijan
https://orcid.org/0009-0007-5744-0591

Latafat Gardashova

Doctor of Technical Sciences, Vice-Rector for Scientific Affairs
Azerbaijan State Oil and Industry University
AZ1010, 20 Azadliq Ave., Baku, Azerbaijan
https://orcid.org/0000-0003-3227-2521

Abstract. The prompt and accurate detection of microdefects in industrial materials is a priority for improving product quality, production safety, and process optimisation. The purpose of this study was to create an automated inspection system that uses artificial intelligence to identify microdefects in industrial materials. The study was conducted on laboratory and industrial samples with microdefects using a multi-sensor system consisting of visual cameras, ultrasound, thermography, and X-rays. The data was pre-processed by filtering, normalising, and extracting contours and analysed using Convolutional Neural Network (CNN), Vision Transformer, and 3D CNN deep learning models with multimodal integration, transfer learning, augmentation, and weight optimisation, with the system performance evaluated by accuracy, precision, recall, and F1-score metrics. A comprehensive analysis showed that the individual use of visual cameras with an accuracy of 92.3%, ultrasonic sensors with an accuracy of 89.5%, thermography with an accuracy of 85.1%, and an X-ray scanner with an accuracy of 95.6% provided high results, and their combination increased the integrated index to 97.8%, which confirms the advantages of the multichannel approach. The use of pre-processing methods (Gaussian and median filters, normalisation, histogram alignment) and augmentation increased the accuracy to 94.1% and the F1-score to 92.6% (compared to the initial 85.2%), while transfer learning increased accuracy by 12-15% and reduced training time, reducing the number of false positives. The system maintained an accuracy of over 90% in noise and variations in production conditions, and at least 80% in extreme scenarios. Practical tests on a server with NVIDIA A6000 GPUs showed an average sample processing time of 120-180 ms (5-8 FPS) and linear scalability with the number of GPUs, which confirmed the system's suitability for integration into real-time industrial systems. The findings of this study can be used by quality control specialists and developers of industrial information and measurement systems to improve the accuracy and efficiency of microdefect detection

Keywords: neural networks; deep learning; computer vision; transfer learning; data augmentation

Introduction

The prompt detection of microdefects in industrial materials is a priority to ensure product quality, equipment safety, and reduce production losses. Conventional non-destructive testing methods have limited sensitivity and depend on the human factor. In this regard, the integration of artificial intelligence (AI) into information and measurement systems opens new opportunities for automated and highly accurate detection of hidden defects. The use of deep learning, computer vision, and multimodal signal processing enables the creation of adaptive control systems that meet the challenges of Industry 4.0.

The integration of artificial intelligence into automated measurement and data evaluation systems has been actively developing in the 21st century. A comprehensive

Suggested Citation:
Mikayilov, K., & Gardashova, L. (2025). Methods of signal processing and data interpretation for detecting microdefects in industrial materials. *Information Technologies and Computer Engineering*, 22(3), 113-124. doi: 10.31649/vitce/3.2025.113

*Corresponding author

analysis of the publications by E. Yusubov (2023), and E. Mustafayev & R. Azimov (2024) revealed that Azerbaijani scientists pay special attention to the practical integration of AI algorithms into specific application systems. Both researchers pointed out that the use of artificial intelligence increases the accuracy, reliability, and productivity of data processing, regardless of the subject area – energy or educational technologies. Additionally, the researchers noted the advantages of automation in reducing the human factor and the ability to scale solutions, which underlines the applied potential of artificial intelligence for local Azerbaijani developments.

Numerous international studies consider different approaches to integrating artificial intelligence into industrial materials quality control and defect detection systems. S. Alamuru *et al.* (2024) conducted a systematic review of deep learning algorithms, including Convolutional Neural Network (CNN) and hybrid models that integrate image and sensor signal analysis. The researchers emphasised that the use of deep neural networks increases the accuracy and speed of microdefect identification and noted that automation of the inspection process reduces the influence of the human factor. A. Deepak & P. Rao (2024) investigated the use of material deformation sensors in combination with artificial intelligence algorithms to detect defects in real time. The researchers detailed the structure of multi-sensor platforms that allow synchronous collection of data on stresses and strains of materials and analyse them using machine learning models. A. Dubaish & A. Jaber (2023) focused on modern signal processing methods and artificial intelligence-based approaches for diagnosing gearbox defects. Their publication analysed digital signal processing, including spectral analysis, wavelet transform, and noise reduction methods, combined with machine learning models.

Analysing the studies by of A. Ercetin *et al.* (2024), Z. Mahamud *et al.* (2025), and A. Saberironaghi *et al.* (2023) in the context of integrating artificial intelligence for quality control of industrial products, several shared trends and differences can be noted. Specifically, all studies demonstrated an increase in the accuracy and productivity of control systems due to automation and the use of AI. A. Ercetin *et al.* (2024) addressed the conventional manufacturing processes, assessing the condition of the surface and tools in machining, and showed that the use of computer vision and machine learning increases accuracy and productivity while minimising the impact of the human factor. Z. Mahamud *et al.* (2025) expanded the focus on additive manufacturing to include energy and biomedical applications and showed that the integration of multimodal sensors with deep neural networks enables the analysis of complex geometries and materials with high accuracy, highlighting the potential of these systems for industrial scale and Industry 4.0 standards. In comparison, the review by A. Saberironaghi *et al.* (2023) covered a wide range of deep learning methods, including CNNs, R-CNNs, and transformers, with a focus on image and video processing. The researchers emphasised that the combination of different deep learning architectures and computer vision algorithms provides the greatest accuracy and efficiency of automated inspection, especially for complex product geometries.

A review of modern machine vision methods for defect detection was performed by Z. Ren *et al.* (2022) and X. Zheng *et al.* (2021), who highlighted the latest advances in the application of deep learning to analyse surface defects in industrial products. The researchers emphasised that the key problems continue to be the limited and unbalanced datasets, the challenge of generalising models to varied materials and textures, and the decrease in accuracy when working with low-quality or noisy images. This indicates that, despite the considerable advancement of CNN and Vision Transformers, the reliability of defect detection systems in industrial environments is still not guaranteed. H.S. Kim *et al.* (2023) studied the non-destructive detection of thin microdefects in FRP (Fibre-reinforced plastic) composites using terahertz electromagnetic waves and CNN. The researchers demonstrated that a combination of sensor technologies and deep learning can effectively detect defects that are difficult to identify using conventional methods.

Despite a significant body of research, most studies focus either on specific types of materials (casting, additive manufacturing, composites) or concrete sensor technologies and rarely cover the integration of multimodal signals. Furthermore, there is a gap in the study of systems that combine various signal processing techniques with deep neural networks for comprehensive automated inspection of microdefects in industrial materials of various types. The purpose of the present study was to develop an artificial intelligence-based information and measurement system for detecting microdefects in industrial materials. The purpose entailed the following tasks: to analyse existing methods of signal and sensor data processing for microdefect detection and determine their advantages and limitations in the context of different types of industrial materials; to develop an experimental strategy for testing the control system, including the selection of sensors and algorithms, to assess the accuracy, reliability, and adaptability of automated defect detection.

Materials and Methods

The study was conducted from June 2023 to July 2025 in the laboratories of the Azerbaijan Institute of Materials Science and at industrial enterprises in Azerbaijan, where 450 samples of metal and composite materials with microdefects were tested, combining controlled experiments and analysis of real production conditions. Primary data was collected using a multi-sensor system that included Basler (Germany) visual cameras, Olympus (Japan) ultrasonic sensors, FLIR (USA) thermal cameras and GE (USA) digital X-ray scanner, with each sample passing through all sensors to obtain a complete set of signals. The collected data underwent pre-processing, including filtering, normalisation, augmentation, and contour extraction, after which deep learning models were used, including CNN, Vision Transformer (ViT), and 3D CNN. CNN and ViT were used to detect, localise, and classify defects, as well as multimodal integration of data from various sensors, which improved the accuracy and reliability of the system.

To solve the problems of detecting, localising, and classifying micro-defects in industrial materials, a step-by-step approach was applied that combined signal pre-processing, deep learning, and multimodal analysis methods. The first stage involved pre-processing the data, including filtering, normalisation, and contour extraction. At the second stage, artificial intelligence models, such as CNN and ViT, played a key role in the processing, ensuring high quality recognition of even subtle microdefects. A 3D CNN was used to process the three-dimensional images obtained with a digital X-ray scanner, which allowed considering the depth of the material structure and detecting hidden internal defects in the sample thickness.

The models implemented using TensorFlow and PyTorch include a CNN for ultrasound signal processing, a Vision Transformer for 2D images, and a 3D CNN for 3D X-ray scans, with OpenCV used for image pre-processing, contour extraction, and visualisation of the results. Data augmentation techniques were employed to train the models, including random rotations, scaling, lighting changes, and mirroring. This helped to improve the generalisability of the models and increase their resilience to the variability of conditions in which images or signals can be obtained in a real industrial environment. The class balancing features were to compensate for the imbalance between defective and healthy samples, where defective samples accounted for about 30% of the total sample and healthy samples for 70%, and oversampling methods were used for defective classes and stratified batching to maintain proportions in training batches. Hyperparameter selection included setting the learning rate within 1e-5-1e-3, batch size from 16 to 64, number of epochs from 50 to 150, regularisation coefficients from 1e-6 to 1e-4, and number of neurons in hidden layers from 128 to 1,024, which allowed reaching an optimum balance between convergence speed and avoiding overfitting.

The models were trained using transfer learning for retraining, as well as optimised methods, including quantisation and pruning, to improve performance and reduce computational costs. Performance was evaluated using the accuracy, precision, recall, and F1-score metrics, where accuracy showed the overall accuracy of the model, i.e., the proportion of correctly classified samples among all test data; precision determined the accuracy of classification of a concrete class, i.e., the proportion of correctly predicted defects among all those predicted as defects; recall measured the ability of the model to detect all defects, i.e., the proportion of correctly found defects among all true defects; and F1-score was a harmonic average of precision and recall and allowed assessing the balance between accuracy and completeness, which was especially relevant in case of class imbalance. To increase the accuracy and reliability of defect detection, all signals from visual cameras, ultrasonic sensors, thermal cameras, and an X-ray scanner were integrated using a multi-channel fusion algorithm that combined information from different sensors at the feature level to provide a total assessment of the material condition. To increase the accuracy and stability of the models, a combination of Dice Loss and Focal Loss functions was implemented, which are effective in case of a considerable imbalance between classes, typical for defect segmentation tasks, where the proportion of the defective area may be insignificant compared to the background. The data was filtered with Gaussian and median filters to prepare it for analysis by CNN, Vision Transformer, and 3D-CNN models to improve their accuracy and stability.

Performance testing and model training were performed on a server with an NVIDIA A6000 GPU (USA) with 48 GB of Video Random Access Memory (VRAM), which enabled high-resolution image processing and parallel training of large transformer architectures. An Intel Xeon CPU and 256 GB of Random Access Memory (RAM) enabled data pre-processing, annotation, and training batch generation without performance losses. The models were trained in the JupyterLab environment. All the components of the training pipeline were containerised using Docker, which allowed creating a reproducible environment with fixed versions of libraries, including PyTorch, OpenCV, and Albumentations, and ensure independence from changes in the system environment.

Results

Evaluation of the accuracy of microdefect detection using a multi-sensor system

Evaluation of the precision of microdefect detection is a key step in ensuring the quality of industrial materials, as even small defects can substantially affect the strength, reliability, and durability of products. Critical micro-defects include surface cracks, internal delaminations, micro-inclusions, cavities, and localised material changes that are challenging to detect using standard inspection methods. The study showed that the multi-sensor system effectively detects a wide range of such defects: visual cameras detected surface cracks, ultrasonic sensors detected internal delaminations and hidden cavities, thermography detected areas with localised changes in thermal conductivity, while an X-ray scanner detected micro-inclusions and micro-cavities in deep layers of the material (Table 1).

Table 1. Precision and recall of microdefects detection by various sensors and multi-sensor system

Sensor channel	Types of defects detected	Precision (%)	Recall (%)
Visual cameras	Surface cracks	92.3	88.7
Ultrasonic sensors	Internal stratification, hidden cavities	89.5	91.2
Thermography	Localised changes in thermal conductivity, hidden cracks	85.1	83.4
X-ray scanner	Microinclusions, microcavities	95.6	94.8
Multi-sensor system	All the above defects	97.8	96.9

Source: developed by the authors

The analysis of the effectiveness of the multi-sensor system revealed that the integration of data from high-resolution visual cameras, ultrasonic sensors, infrared thermography, and a digital X-ray scanner can improve the accuracy of microdefect detection compared to the use of individual sensors. According to the test results, the visual cameras demonstrated high efficiency in detecting surface defects with an average Precision of 92.3% and Recall of 88.7%. The main limitation of the optical channel was the dependence of the quality on the lighting and the orientation of the defect relative to the camera. Ultrasonic sensors provided reliable detection of internal delamination and hidden cavities with 89.5% accuracy and 91.2% recall but were less effective for very small surface cracks. The thermographic channel showed an average precision of 85.1% and a recall of 83.4%, proving to be excellent for detecting areas with localised changes in thermal conductivity, including cracks hidden under the surface that have not yet reached a critical size. The X-ray scanner demonstrated the greatest accuracy in detecting micro-inclusions and micro-cavities in the inner layers of the material – 95.6% in precision and 94.8% in recall.

The integration of all sensor data using the multi-channel fusion algorithm resulted in a total precision of 97.8% and a recall of 96.9%, which is 4.5-9.2% greater than the results of any individual channel. An analysis of misclassification examples showed that the multi-sensor approach compensated for the weaknesses of each individual sensor: cases where the visual camera missed a crack due to a blip or shadow were correctly processed by the X-ray channel or ultrasonic analysis; defects that were not visible on X-rays due to low contrast were detected by thermography or optical analysis. The results are presented in Figure 1 for ease of comparison.

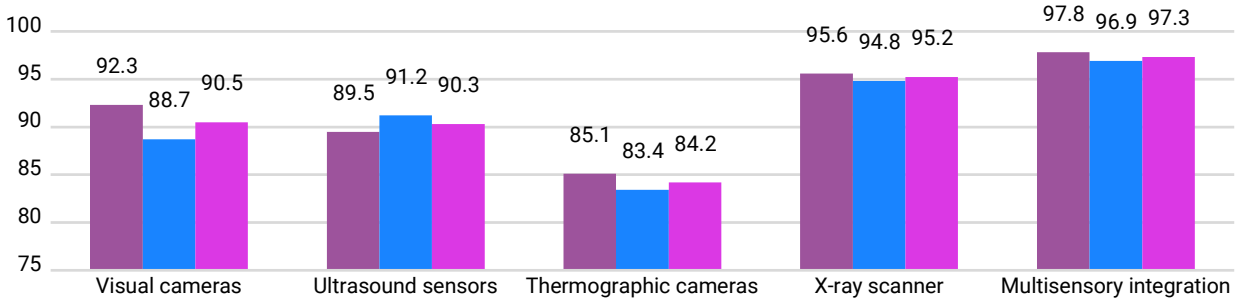


Figure 1. Comparative accuracy of micro defect detection by different sensors and integrated multi-sensor system
Source: developed by the authors

Figure 1 demonstrates that the individual sensors have distinct specialisations: visual cameras demonstrate high precision and F1-score for surface defects; ultrasonic sensors are better at detecting internal delamination and cavities, showing balanced Precision and Recall; thermographic cameras are effective at capturing hidden defects such as localised changes in thermal conductivity; while the X-ray scanner provides the greatest accuracy and completeness for micro-inclusions and micro-cavities in the inner layers of the material. Multi-sensor integration combines the advantages of all channels to achieve maximum accuracy and F1-score.

Comparison of the effectiveness of deep learning models for processing different types of data

Comparison of the effectiveness of deep learning models was a key step in optimising the process of detecting, localising, and classifying micro-defects in industrial materials, as different types of input data and defect specifics required the use of models with different architectural features. This analysis helped to determine which algorithms provided the greatest precision, recall, and balance of results in real-world production conditions, and helped reduce the risk of missing critical defects.

CNN worked effectively with 2D images and ultrasonic signals, providing high precision in local defect detection due to its ability to highlight local patterns. ViT had a better ability to adapt to the variability of lighting conditions and surface textures, which increased the generalisation ability when processing 2D images, although the processing time increased. 3D-CNN specialised in volumetric data, providing the highest efficiency in detecting internal microdamages and balanced accuracy and completeness. ViT-3D combined the properties of transformers with three-dimensional computing, increasing the completeness of defect detection in 3D scans, but was slightly inferior to 3D-CNN in terms of accuracy and precision.

For 2D images, transformers proved to be more accurate than convolutional networks, demonstrating a better balance between precision and recall and better robustness to changes in illumination and texture, although this required longer processing times. In ultrasound data analysis, CNNs were more accurate overall, but ViTs showed a better ability to detect defects in noisy signals, reducing the risk of misses. For volumetric X-ray scans, 3D CNNs provided the greatest performance and the best combination of accuracy and completeness, while ViT-3D had slightly lower overall performance but with greater sensitivity to hidden damage. The combined use of the models confirmed the feasibility of the hybrid approach: ViTs are best for detecting surface defects, while 3D-CNNs are best for internal defects, which together increases the overall efficiency of the system (Table 2).

Table 2. Comparison of the effectiveness of deep learning models for different types of data

Data type	Model	Accuracy	Recall	Precision	F1-score
2D images	CNN	0.93	0.91	0.94	0.925
	ViT	0.95	0.94	0.96	0.95
Ultrasound data	CNN	0.89	0.92	0.87	0.895
	ViT	0.88	0.94	0.85	0.89
Volumetric X-ray	3D-CNN	0.96	0.95	0.97	0.96
	ViT-3D	0.94	0.96	0.92	0.94

Source: compiled by the authors

Overall, the results showed that the choice of model depended on the type of input data and the particular task: for 2D images, the identification of surface defects was more effective with Vision Transformer, for ultrasound signals – CNN under the condition of high data purity, and for 3D scans – 3D-CNN as the most balanced solution for internal defects. In further research, each model was applied according to the type of data and the specifics of the defects, i.e., all models were not used simultaneously on the same data set, but the optimal model was selected for each sensor channel. Comparison of the results of different models helped to determine their effectiveness and reasonably combine them in a hybrid architecture to improve the overall accuracy of the system and reduce the risk of missing defects of different nature.

Effect of signal pre-processing and augmentation methods on the quality of defect detection

Evaluation of the impact of signal pre-processing and augmentation methods on the quality of microflaw detection was critical to improving system reliability. Effective noise removal, signal normalisation, illumination correction, and expansion of the training dataset improved the accuracy and stability of deep learning models, reducing the risk of defect misses and overtraining. The analysis of the implemented methods revealed that their integrated application substantially improved the main quality metrics of the models, which was of practical significance for implementation in production and ensuring high reliability of microflaw detection.

A comparison of the accuracy of microflaw detection by deep learning models on data using a Gaussian filter and a median filter showed an average increase of 4-6%, which reflected the effectiveness of using these pre-processing methods to reduce noise and improve signal quality before training models. The use of normalisation and histogram equalisation increased the average F1-score of the models by 3-5%. Data augmentation increased the training dataset fivefold, which markedly reduced the risk of overtraining and increased recall and F1-score by 6-9%. The integrated application of noise filtering, normalisation, histogram alignment, and augmentation resulted in a 94% increase in the accuracy of microdefect detection, which is substantially greater than the initial results. To illustrate the impact of different pre-processing and augmentation methods, Table 3 presents the key model quality metrics (accuracy, precision, recall, F1-score) depending on the approach used. The table clearly shows a gradual improvement in the results: starting from the work without processing (accuracy 85.2%) and achieving the best results when all methods are combined (accuracy 94.1%, F1-score 92.6%).

Table 3. Comparison of the accuracy, precision, recall, and F1-score of microflaw detection models depending on the data processing method

Processing method	Accuracy, %	Precision, %	Recall, %	F1-score, %
No preprocessing	85.2	83.9	82.5	83.2
Gaussian filter	88.1	86.5	85.7	86.1
Median filter	87.5	85.9	85.1	85.5
Gaussian + median filter	89.3	87.8	87.2	87.5
Normalisation + histogram alignment	91.0	89.6	89.1	89.3
Augmentation	92.4	91.2	90.8	91.0
All methods together	94.1	92.8	92.5	92.6

Source: compiled by the authors

The logic of data pre-processing and augmentation was summarised in the form of a sequential diagram, presented in Figure 2. Firstly, the raw data was passed through a noise filtering stage using Gaussian and median filters, which reduced random artefacts and improved the signal quality for further processing. The next step was normalisation, which unified the intensity scale of pixels or signals, reducing the impact of variations in lighting and sensor settings. Subsequently, the data were subjected to histogram equalisation, which ensured an even distribution of brightness and improved contrast, making it easier to detect faint defects. Finally, augmentation techniques such as rotation, scaling, brightness changes, and mirroring were applied to increase the diversity of the training set, reduce the risk of overfitting, and improve the stability of the models when processing new data.

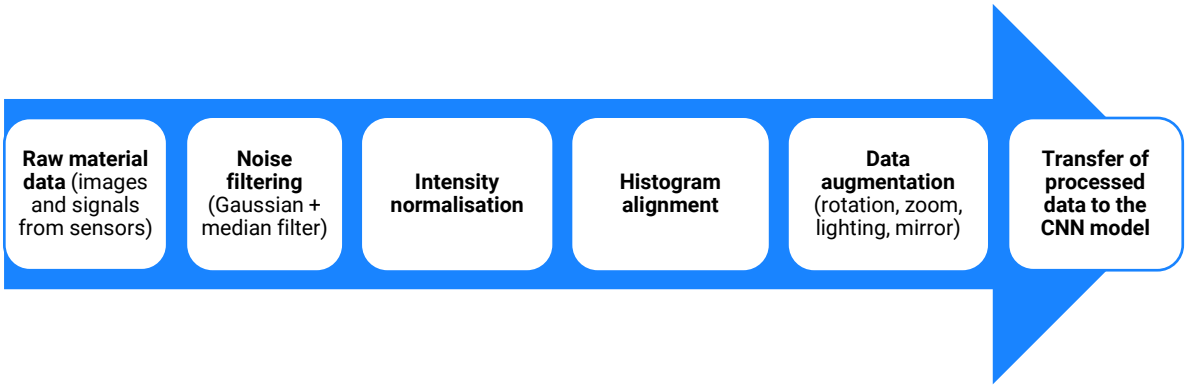


Figure 2. Data pre-processing and augmentation pipeline
Source: created by the authors

The processed data were used as input parameters for deep learning models to classify and segment defects. Such a structured sequence of actions enables maximum preparation of the input data, reducing the impact of noise and variations, which improves the efficiency of the microflaw detection system. Thus, the integrated use of pre-processing and augmentation methods is a key factor in improving the accuracy and reliability of depth models in industrial materials quality control tasks, making them suitable for practical implementation in real production environments.

Analysis of the results of model retraining using transfer learning

The results of the study clearly demonstrated the advantages of retraining models using transfer learning compared to learning from scratch. The same model architecture was employed for the analysis, which underwent two training modes: training from scratch, where all weights were initialised randomly and the model was trained only on its own dataset, and retraining using transfer learning, when the model adapted general features from previously trained models to the specifics of local microdefects. The models trained from scratch achieved an average accuracy of 81.7%, precision of 79.5%, recall of 78.3%, and F1-score of 78.9%. At the same time, the same architectures trained using transfer learning achieved an accuracy of 93.1%, precision of 91.8%, recall of 90.5%, and F1-score of 91.1%, which corresponds to an increase in accuracy of approximately 12-15% for all key metrics. This increase was explained by the fact that transfer learning allowed the model to use already formed general features – contours, textures, and intensity gradations – and adapt them to the specifics of local microdefects. This approach proved to be particularly effective for defects with low contrast, heterogeneous structure, or rare distribution, where the proprietary dataset contained a limited number of annotated examples. For example, when detecting small internal inclusions or microcavities, transfer learning increased recall by 12% and F1-score by 12.2% compared to learning from scratch, which meant fewer missed defects and a reduction in the number of false positives. For a visual comparison, Figure 3 presents the key metrics of the same model in the two learning modes.

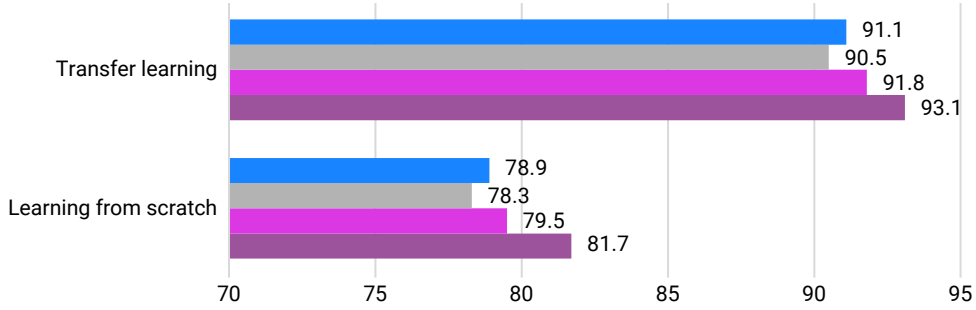


Figure 3. Comparison of key metrics of microflaw detection models: learning from scratch and transfer learning
Source: compiled by the authors

The models using transfer learning demonstrated an improvement in all metrics compared to models that were trained from scratch. This indicates that pre-training on large and diverse data generates useful features that enable better recognition of subtle and complex microflaw patterns, even with a limited number of specific training examples. As a result, transfer learning not only improves accuracy, but also reduces training time and the need for large amounts of annotated data, which is essential for the practical implementation of quality control systems in industrial settings.

The analysis also revealed that the training sample size has a different impact on both approaches. Training from

scratch requires more data to achieve stable results, while transfer learning models perform well even on limited data sets, and the increase in metrics is slower with increasing sample size, reflecting their ability to quickly generalise information. Additionally, transfer learning reduces training time and resource costs: models were trained on local samples faster than with full training from scratch, which provides a practical advantage for regularly updating models in a production environment. This approach allows maintaining recognition accuracy even in the face of variations in lighting, noise, and other real-world production conditions. Thus, the use of transfer learning in combination with carefully selected model architectures allows optimising the learning process and improving the results of micro defect recognition, making this approach promising for practical implementation in automated quality control systems in industry.

Resistance of models to noise and variations in production conditions

The robustness of the models to noise and variations in production conditions is key to the practical application of the system in real-world production, where data often contains artefacts, variable lighting, and electronic interference. High robustness ensures reliable detection of micro-defects even in the presence of noise or partial overlap of defective zones, minimises the number of missed defects and false positives, and guarantees stable system performance in different conditions. The robustness of the models to noise and variations in production conditions was assessed by experimentally simulating random interference, lighting changes, and equipment artefacts on the input data, which helped to verify their practical applicability. The evaluation results showed significant patterns. The models performed best on clean data, achieving a high balance between accuracy and recall and correctly identifying most micro-defects. The addition of noise led to a moderate drop in all metrics, reflecting an increase in the number of false positives and false negatives. Changes in lighting primarily affected the recall metric, increasing the proportion of missed defects due to reduced contrast. Shadows became an even more serious challenge, as they overlapped the defective areas and caused a more significant decrease in performance. The most challenging conditions were glare on glossy and metallic surfaces, where the number of false positives increased dramatically and the localisation of damage became more complicated. Overall, the models demonstrated robustness in standard noise conditions but lost up to 10-12% of their performance in extreme scenarios, which underscores the need for adaptive processing methods and multisensory integration. The largest drop in recall reflects an increased probability of missing rare or weak defects, while the decrease in precision is caused by false positives for noise artefacts. The use of multimodal integration, which combines visual, ultrasonic, thermal, and X-ray data, compensates for the limitations of individual sensors and maintains accuracy above 80% even in challenging conditions. To illustrate the impact of different types of noise and variations in production conditions on the quality of the models, Figure 4 presents the key quality metrics (accuracy, precision, recall, and F1-score) in different experimental scenarios.

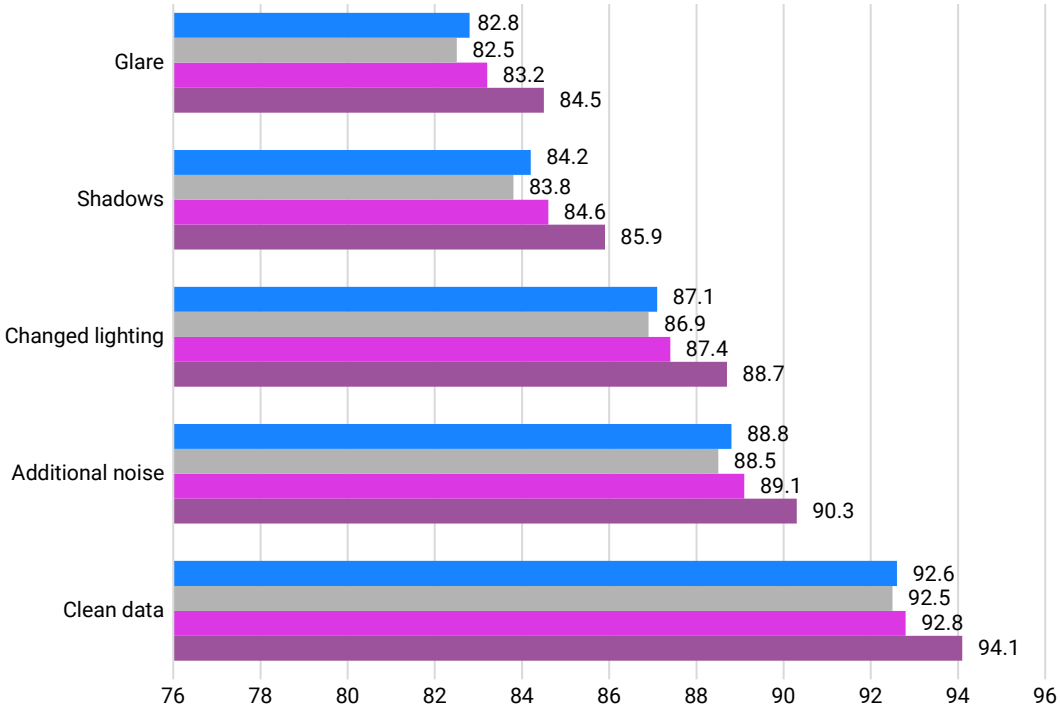


Figure 4. Resistance of models to noise and variations in production conditions
Source: compiled by the authors

As Figure 4 demonstrates, the presence of noise and lighting variations leads to a gradual decline in model quality. The influence of shadows and glare is significant, often leading to false positives and false negatives. At the same time, the model demonstrated a fairly prominent level of stability, maintaining an accuracy of over 80% even in the most challenging conditions. Overall, the analysis revealed that the developed models are highly ready to work in conditions of production noise and variations, with clearly defined ways to further improve the system to minimise false positives and false negatives, which is critical for implementation in automated quality control systems.

Practical efficiency and performance of the model in the computing environment

The average processing time per sample for both models ranged within 120-180 milliseconds, which ensured processing at approximately 5-8 frames per second (FPS), in line with the requirements of real-time systems in industrial flaw detection. The type and size of the input data affected the processing time: CNNs exhibited faster processing due to fewer parameters, while ViT transform architectures required additional time due to the global attention mechanism and more complex computational structure. In terms of resources, both models worked efficiently within the available hardware: the maximum peak VRAM consumption did not exceed 38 GB even when simultaneously processing battles from multiple samples, and RAM was used up to 256 GB, which avoided server overload. Optimisation of models through quantisation and pruning further reduced resource consumption without significant loss of accuracy, increasing scaling efficiency. The scalability of the system was assessed as a linear increase in performance with the number of GPUs, making the training and inference process flexible and enabling seamless upscaling for larger data volumes. The use of the high-performance NVIDIA A6000 GPU ensured stable operation of both models at high image resolution, which is critical for accurate detection of small micro-defects. The deployment platforms included both on-premises servers and cloud-based solutions with Docker containerisation, which ensured reproducibility and flexibility when integrating the models into production processes. Key indicators of the practical effectiveness of both models are presented in Table 4.

Table 4. Performance and resource characteristics of the computing environment for micro defect detection models

Indicator	Value
Average sample processing time	120-180 ms
Maximum VRAM consumption	Up to 38 GB
RAM	Up to 256 GB
GPU usage	NVIDIA A6000
Performance (FPS)	5-8 FPS
Scalability	Linear growth with GPU count
Deployment platforms	Local servers, cloud (Docker)

Source: compiled by the authors

Thus, the evaluations confirmed that the proposed approach to microdefect detection is highly effective when using off-the-shelf deep learning models, demonstrating simultaneous accuracy, speed, and resource efficiency. This indicates the practical applicability of the methodology for integration into production quality control systems. Overall, the study findings confirmed that a comprehensive combination of multimodal data integration and adaptive preprocessing can markedly improve the reliability and stability of detecting various microdefects. Furthermore, the proposed approach is versatile and can be scaled to different production environments and types of materials without significant model redesign.

Discussion

The obtained findings are in line with the trends recorded in the 2021-2025 studies, which also demonstrate the active development of methods for using big data and artificial intelligence to monitor and assess the condition of technical objects in various industries. Specifically, in the field of additive manufacturing of metal parts, particularly those made by selective laser melting, there is a tendency to integrate monitoring data into damage models to predict the technical condition of metal parts. X. Yan & H. Fu (2024) considered a method for predicting the technical condition of parts manufactured by selective laser melting based on big data analysis and artificial intelligence. The researchers proposed a combined model that integrates sensor monitoring data with information about internal material defects, which allows predicting wear and the probability of parts failure. X. Yan & H. Fu focused on predicting the condition of the material during operation, while the present study focused on the combined assessment of the accuracy of microdefect detection in different sensor channels. This explains why a direct correlation of accuracy results between studies is not always observed.

In the field of non-destructive evaluation of the aging of insulating coatings, M. Qiu & X. Ge (2025) proposed a method that combines ultrasonic guided waves with signal processing and machine learning methods to evaluate the aging of coatings. The researchers emphasised the significance of signal pre-processing to increase the sensitivity of the algorithms to small changes in the coating structure. M. Qiu & X. Ge focused more on the spectral characteristics of the signal, while the present analysis integrated results from several types of sensors for comprehensive

defect assessment. This explains the difference in accuracy and types of defects detected.

In the field of automated optical inspection and quality assessment of semiconductor and optoelectronic products, A. Ebayyeh & M. Abd Al Rahman (2022) focused on the use of CNNs for automatic defect detection and quality assessment, which enables increased efficiency and accuracy of quality control processes. Compared to the present study, the researchers' approach was more narrowly focused on a concrete type of product. S. Jiao *et al.* (2025) proposed an approach to detecting microdefects on semiconductor wafers that combines optical topography with a lightweight classification network for real-time data processing. The researchers emphasised the advantages of a multimodal approach, which enables faster and more accurate detection. Compared to the current study, the cited study demonstrated the potential of integrating optical data with a lightweight classification network. At the same time, the current study extended the concept of multimodal integration by combining optical, ultrasonic, and thermal sensors, which resulted in slightly greater overall accuracy in a laboratory setting.

J.L. Tai *et al.* (2025) performed a systematic review of modern methods of non-destructive testing of composite materials and successful practical implementations, noting that the combined use of ultrasonic, visual, and electromagnetic technologies markedly increases the accuracy and reliability of defect detection. Z. Zhao (2021) reviewed non-destructive testing methods for the detection of defects in ceramic materials, including ultrasonic, X-ray, and infrared technologies, emphasising that the accuracy and reliability of detection strongly depend on the choice of sensor technique and signal pre-processing. The current approach confirmed these findings, while demonstrating the additional integration of machine learning algorithms for defect classification, which provided increased efficiency and versatility in detecting various types of microdefects in different materials.

A. Mohammed & M. Hussain (2025) conducted a technical review of the use of deep learning for automated welding defect detection. The researchers emphasised the significance of data preparation and CNN and Transformer architectures for the accuracy and reliability of detection. Compared to the present study, the authors focused on a concrete production task of welding. D. Na *et al.* (2025) investigated the detection of microdefects in multilayer ceramic capacitors based on the instantaneous frequency of the electromechanical response. The researchers showed that analysing the frequency response of the signal can increase the sensitivity to small defects. The approach of D. Na *et al.* was focused on one specific type of sensor and material, while the present study combined data from several sensors to provide a more comprehensive assessment of microdefects. Z. He *et al.* (2025) developed a new system for visualising microdefects on highly reflective surfaces by combining optical technologies with image processing algorithms. The researchers noted that their system allows increasing the contrast of defects and reducing the influence of reflections. The method employed in the present study involved the integrated processing of data from different sensors, which provided a more complete assessment of defects in various materials and lighting conditions.

Q. Xiao *et al.* (2023) proposed a method for detecting defects in transparent components based on a modified YOLOv7 algorithm. The researchers emphasised that the increased accuracy and speed are achieved by optimising the architecture and adding mechanisms for localising small defects. They focused on a single type of material and use only deep neural networks for imaging, while the presented approach employed multi-sensor analysis. D. Ashebir *et al.* (2024) discussed the challenges and progress in the detection of multi-scale defects in the additive manufacturing of thermoplastic fibre-reinforced composites. The researchers emphasised the need for a combination of multiple NDT techniques and algorithms to accurately detect defects at different scales. The presented study correlates with the findings of D. Ashebir *et al.* and confirmed the significance of the multi-sensor approach but demonstrated its effectiveness in a wider range of materials and defect types, which extends its practical application.

Eddy current and tunnelling magnetoresistance sensors are considered as effective approaches to detecting microdefects in metal materials. K.S. Tran *et al.* (2024) emphasised that the approach effectively evaluates both surface and subsurface defects but was limited to steel filaments and a single sensor. B. Wei *et al.* (2025) showed high sensitivity to small defects in aluminium foil by compensating for the spatial position of the differential sensor, but the method also had limitations for transparent or reflective surfaces. In contrast to these approaches, the current multi-sensor method provided a wider range of defect inspection through combined data acquisition and additional signal processing, enabling effective defect detection in a variety of materials and complex manufacturing environments.

K.D. Malakar *et al.* (2025) analysed digital image processing for monitoring complex structures, including contrast enhancement and defect detection. The researchers emphasised the significance of data pre-processing, such as normalisation and histogram alignment, which helps to improve the accuracy of the algorithms. These findings correlate with the observations of the present study: the use of pre-processing and augmentation increased the F1-score of models by 3-5%, confirming the effectiveness of such methods for stabilising models under variations in lighting and noise. N. Prottasha *et al.* (2022) considered the use of BERT-based (Bidirectional Encoder Representations from Transformers) transfer learning for text data analysis tasks. The researchers showed that retraining of pre-trained models can markedly improve accuracy and F1-score even on limited data sets. These findings are fully correlated with the results obtained in the present study, where retraining ViT-based models increased accuracy by 12-15% compared to training from scratch, which reflects the effectiveness of transfer learning for adapting models

to specific defects and limited samples. Thus, the aforementioned studies confirmed that the key success factors are a combination of pre-processing, augmentation, and transfer learning, which was confirmed by the findings obtained for industrial flaw detection systems.

Comparison of the findings with published studies demonstrated that the application of a multi-sensor approach combined with data pre-processing, augmentation, and model retraining using transfer learning led to greater accuracy and versatility in detecting microflaws in various materials and production conditions. Specifically, the integration of data from multiple sensors provided a more comprehensive defect assessment compared to methods focused on individual sensors or specific materials, as evidenced by the steady growth of key metrics such as accuracy, recall, and F1-score. In some cases, there were slight discrepancies with the findings of other researchers, e.g., in processing speed or sensitivity to concrete types of defects, which can be explained by the more complex multi-sensor architecture and variations in experimental conditions. Overall, the conducted study confirmed the effectiveness of the integrated approach and at the same time expanded the scope of practical application to a wider range of materials and defects, demonstrating consistency with the general trends in the development of automated defect detection methods.

Conclusions

The study found that the developed microdefect detection system proved to be efficient and reliable, meeting the modern requirements of industrial quality control automation. The multi-sensor system improves the accuracy of microdefect detection compared to individual sensors. Visual cameras achieved an accuracy rate of 92.3%, ultrasound sensors – 89.5%, thermography – 85.1%, and an X-ray scanner – 95.6%. The combination of all sensors increased the accuracy to 97.8%, confirming the benefits of a multi-channel approach. Vision Transformer showed the best results for 2D images (95% accuracy), CNN for ultrasound data (89% accuracy), and 3D-CNN for 3D X-ray scans (96% accuracy). The combination of models based on the data type helped to maximise the quality of defect detection. Gaussian and median filter signal pre-processing methods increased the accuracy of defect detection by 4-6%. Normalisation and histogram alignment added 3-5% to the improvement, while augmentation increased recall and F1-score by 6-9%. The integrated application of all methods resulted in an increase in accuracy to 94.1% and F1-score to 92.6%, which exceeds the initial level (85.2%). This demonstrates the critical role of complex pre-processing in improving the quality of models in real-world production environments.

The use of transfer learning to retrain the models increased recognition accuracy by 12-15% compared to training from scratch. This approach is especially effective with a limited amount of annotated data and complex defects with low contrast. Furthermore, transfer learning reduced the training time and reduced the number of false positives and false negatives, increasing the reliability and resource efficiency of the system. The models under consideration demonstrated high resistance to noise and variations in production conditions, maintaining an accuracy of over 90% when using complex pre-processing and augmentation. Under extreme conditions, the accuracy decreased but did not fall below 80%. The proposed additional methods of filtering, lighting correction, and multimodal analysis effectively minimised false positives, which is essential for implementation in production. Practical evaluation of the models on a server with NVIDIA A6000 GPUs showed that the average processing time for one sample was 120-180 ms, while performance reached 5-8 FPS. The maximum GPU memory consumption did not exceed 38 GB. The system demonstrated linear scalability with GPU count and compatibility with industrial platforms, making it suitable for integration into real-time systems. The system's effectiveness was limited by the quality of the training data, the complexity of processing under extreme conditions, and the high demands on computing resources. Further research should focus on developing adaptive models, expanding multimodal integration, and optimising for edge devices with automatic error correction.

Acknowledgements

None.

Funding

The study was not funded.

Conflict of Interest

None.

References

[1] Alamuru, S., Reddy, G.S., & Raju, M.J. (2024). Artificial intelligence and machine learning for defect detection in castings. *Journal of Physics: Conference Series*, 2837, article number 012079. doi: 10.1088/1742-6596/2837/1/012079.

[2] Ashebir, D.A., Hendlmeier, A., Dunn, M., Arablouei, R., Lomov, S.V., Di Pietro, A., & Nikzad, M. (2024). Detecting multi-scale defects in material extrusion additive manufacturing of fiber-reinforced thermoplastic composites: A review of challenges and advanced non-destructive testing techniques. *Polymers*, 16(21), article number 2986. doi: 10.3390/polym16212986.

[3] Deepak, A., & Rao, P. (2024). Real time defect identification using advanced artificial intelligence based material strain sensors for environmental safety. *Journal of Electrical Systems*, 20(1), 33-40. doi: 10.52783/jes.660.

[4] Dubaish, A.A., & Jaber, A.A. (2023). State-of-the-art review into signal processing and artificial intelligence-based approaches applied in gearbox defect diagnosis. *Engineering and Technology Journal*, 42(1), 157-172. doi: 10.30684/etj.2023.142462.1535.

[5] Ebayyeh, A., & Abd Al Rahman, M. (2022). *Deep learning for automatic optical inspection and quality evaluation of semiconductor and optoelectronic manufacturing*. London: Brunel University London.
[6] Ercetin, A., Der, O., Akkoyun, F., Gowdru Chandrashekarappa, M.P., Şener, R., Çalışan, M., Olgun, N., Chate, G., & Bharath, K.N. (2024). Review of image processing methods for surface and tool condition assessments in machining. *Journal of Manufacturing and Materials Processing*, 8(6), article number 244. doi: 10.3390/jmmp8060244.
[7] He, Z., Lin, S., Xiao, Y., Fang, H., & Sun, L. (2025). A novel micro-defects imaging system for high-reflective material surfaces. *IEEE Transactions on Instrumentation and Measurement*, 74, article number 4501814. doi: 10.1109/TIM.2025.3541809.
[8] Jiao, S., Yang, W., Wu, C., Li, Y., & Xue, B. (2025). Mixed-type micro-defect detection in semiconductor wafers: A dual-modal feature real-time detection approach via optical topography and lightweight classification network. *Engineering Applications of Artificial Intelligence*, 160, article number 111838. doi: 10.1016/j.engappai.2025.111838.
[9] Kim, H.S., Park, D.W., Kim, S.I., Oh, G.H., & Kim, H.S. (2023). Non-destructive detection of thin micro-defects in glass reinforced polymer composites using a terahertz electro-magnetic wave based on a convolution neural network. *Composites Part B: Engineering*, 257, article number 110694. doi: 10.1016/j.compositesb.2023.110694.
[10] Mahamud, Z.H., Khan, M.R., Amin, J.M., & Islam, M.S. (2025). AI for defect detection in additive manufacturing: Applications in renewable energy and biomedical engineering. *Strategic Data Management and Innovation*, 2(1), 1-20. doi: 10.71292/SDMI.V2I01.8.
[11] Malakar, K.D., Roy, S., & Kumar, M. (2025). Digital imaging: Processing and analysis. In K. Das Malakar, S. Roy & M. Kumar (Eds.), *Geospatial technologies in coastal ecologies monitoring and management* (pp. 257-289). Cham: Springer. doi: 10.1007/978-3-031-92017-2_8.
[12] Mohammed, A., & Hussain, M. (2025). Advances and challenges in deep learning for automated welding defect detection: A technical survey. *IEEE Access*, 13, 94553-94569. doi: 10.1109/ACCESS.2025.3574083.
[13] Mustafayev, E., & Azimov, R. (2024). Computer system of evaluation of the mass exam results based on recognition of handprinted Azerbaijani characters. In G. Mammadova, T. Aliev & K. Aida-zade (Eds.), *Information technologies and their applications* (pp. 171-183). Cham: Springer. doi: 10.1007/978-3-031-73420-5_15.
[14] Na, D., Choi, M., Yuan, F.G., & Kim, H. (2025). Detection of microdefects in multilayer ceramic capacitors using the instantaneous frequency in the electromechanical response. *Measurement*, 244, article number 116528. doi: 10.1016/j.measurement.2024.116528.
[15] Prottasha, N.J., Sami, A.A., Kowsher, M., Murad, S.A., Bairagi, A.K., Masud, M., & Baz, M. (2022). Transfer learning for sentiment analysis using BERT based supervised fine-tuning. *Sensors*, 22(11), article number 4157. doi: 10.3390/s22114157.
[16] Qiu, M., & Ge, X. (2025). Nondestructive evaluation of aging failure in insulation coatings by ultrasonic guided wave based on signal processing and machine learning. *Coatings*, 15(3), article number 347. doi: 10.3390/coatings15030347.
[17] Ren, Z., Fang, F., Yan, N., & Wu, Y. (2022). State of the art in defect detection based on machine vision. *International Journal of Precision Engineering and Manufacturing – Green Technology*, 9(2), 661-691. doi: 10.1007/s40684-021-00343-6.
[18] Saberironaghi, A., Ren, J., & El-Gindy, M. (2023). Defect detection methods for industrial products using deep learning techniques: A review. *Algorithms*, 16(2), article number 95. doi: 10.3390/a16020095.
[19] Tai, J.L., Sultan, M.T., Łukaszewicz, A., Józwik, J., Oksiuta, Z., & Shahar, F.S. (2025). Recent trends in non-destructive testing approaches for composite materials: A review of successful implementations. *Materials*, 18(13), article number 3146. doi: 10.3390/ma18133146.
[20] Tran, K.S., Shirinzadeh, B., & Smith, J. (2024). Eddy current-based identification and depth investigation of microdefects in steel filaments. *Sensors*, 24(16), article number 5101. doi: 10.3390/s24165101.
[21] Wei, B., Ma, Q., Wen, S., Zhang, J., Peng, L., & Huang, S. (2025). Aluminium foil micro-defect detection method based on motion-induced eddy current using differential TMR sensor spatial position compensation. *Nondestructive Testing and Evaluation*. doi: 10.1080/10589759.2025.2456674.
[22] Xiao, Q., Huang, J., Huang, Z., Li, C., & Xu, J. (2023). Transparent component defect detection method based on improved YOLOv7 algorithm. *International Journal of Pattern Recognition and Artificial Intelligence*, 37(14), article number 2350030. doi: 10.1142/S0218001423500301.
[23] Yan, X., & Fu, H. (2024). Opportunities and challenges for predicting the service status of SLM metal parts under big data and artificial intelligence. *Materials*, 17(22), article number 5648. doi: 10.3390/ma17225648.
[24] Yusubov, E. (2023). Design of an intelligent information measurement system for photovoltaic dc microgrids. *Proceedings of Azerbaijan High Technical Educational Institutions*, 34(11), 54-64. doi: 10.36962/PAHTEI34112023-54.
[25] Zhao, Z. (2021). Review of non-destructive testing methods for defect detection of ceramics. *Ceramics International*, 47(4), 4389-4397. doi: 10.1016/j.ceramint.2020.10.065.
[26] Zheng, X., Zheng, S., Kong, Y., & Chen, J. (2021). Recent advances in surface defect inspection of industrial products using deep learning techniques. *International Journal of Advanced Manufacturing Technology*, 113(1), 35-58. doi: 10.1007/s00170-021-06592-8.

Методи обробки сигналів та інтерпретації даних для виявлення мікродефектів у промислових матеріалах

Канан Мікаїлов

Аспірант
Азербайджанський державний університет нафти і промисловості
AZ1010, просп. Азадлик, 20, м. Баку, Азербайджан
https://orcid.org/0009-0007-5744-0591

Латафат Гардашова

Доктор технічних наук, проректор з наукових питань
Азербайджанський державний університет нафти і промисловості
AZ1010, просп. Азадлик, 20, м. Баку, Азербайджан
https://orcid.org/0000-0003-3227-2521

Анотація. Швидке та точне виявлення мікродефектів у промислових матеріалах є пріоритетом для поліпшення якості продукції, безпеки виробництва та оптимізації процесів. Метою цього дослідження було створення автоматизованої системи контролю, яка використовує штучний інтелект для виявлення мікродефектів у промислових матеріалах. Дослідження проводилося на лабораторних та промислових зразках з мікродефектами за допомогою мультисенсорної системи, що складалася з візуальних камер, ультразвуку, термографії та рентгенівських променів. Дані були попередньо оброблені шляхом фільтрування, нормалізації та вилучення контурів і проаналізовані за допомогою моделей глибокого навчання Convolutional Neural Network (CNN), Vision Transformer та 3D CNN з мультимодальною інтеграцією, трансферним навчанням, аугментацією та оптимізацією ваги, а продуктивність системи оцінювалася за показниками точності, прецизійності, відтворення та F1-балу. Комплексний аналіз показав, що окреме використання візуальних камер з точністю 92,3 %, ультразвукових датчиків з точністю 89,5 %, термографії з точністю 85,1 % та рентгенівського сканера з точністю 95,6 % дало високі результати, а їх поєднання підвищило інтегрований індекс до 97,8 %, що підтверджує переваги багатоканального підходу. Використання методів попередньої обробки (гаусові та медіанні фільтри, нормалізація, вирівнювання гістограми) та аугментації підвищило точність до 94,1 % та F1-показник до 92,6 % (порівняно з початковими 85,2 %), тоді як трансферне навчання підвищило точність на 12–15 % та скоротило час навчання, зменшивши кількість помилкових спрацьовувань. Система підтримувала точність понад 90 % в умовах шуму та коливань виробничих умов і не менше 80 % в екстремальних сценаріях. Практичні випробування на сервері з графічними процесорами NVIDIA A6000 показали середній час обробки зразка 120–180 мс (5–8 FPS) і лінійну масштабованість за кількістю графічних процесорів, що підтвердило придатність системи для інтеграції в промислові системи реального часу. Результати цього дослідження можуть бути використані фахівцями з контролю якості та розробниками промислових інформаційних і вимірювальних систем для підвищення точності та ефективності виявлення мікродефектів

Ключові слова: нейронні мережі; глибоке навчання; комп'ютерний зір; трансферне навчання; збільшення обсягу даних

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 08.07.2025 Revised: 15.10.2025 Accepted: 23.12.2025

UDC 004.8:005.8 DOI: 10.31649/vitce/3.2025.125

Integrated assessment of system privacy: Formalisation, normalisation and differential privacy

Dmytro Prokopovych-Tkachenko*

PhD in Technical Sciences, Associate Professor
University of Customs and Finance
49000, 2/4 Volodymyra Vernadskoho Str., Dnipro, Ukraine
https://orcid.org/0000-0002-6590-3898

Liudmyla Rybalchenko

PhD in Economics, Associate Professor
University of Customs and Finance
49000, 2/4 Volodymyra Vernadskoho Str., Dnipro, Ukraine
https://orcid.org/0000-0003-0413-8296

Volodymyr Zvieriev

PhD in Technical Sciences, Associate Professor
State University of Trade and Economics
02156, 19 Kyoto Str., Kyiv, Ukraine
https://orcid.org/0000-0002-0907-0705

Borys Khrushkov

Postgraduate Student
University of Customs and Finance
49000, 2/4 Volodymyra Vernadskoho Str., Dnipro, Ukraine
https://orcid.org/0009-0002-3978-5012

Valerii Bushkov

Postgraduate Student
State University of Trade and Economics
02156, 19 Kyoto Str., Kyiv, Ukraine
https://orcid.org/0009-0005-5097-2689

Abstract. Requirements for confidentiality and greater data privacy are constantly growing. The aim of this work was to develop a formalised approach to assessing the privacy of information systems based on a vector representation of a set of parameters. In the proposed approach, each parameter has a numerical value within a defined range that reflects the degree of its implementation or importance. For convenience and structure, the parameters were divided into several categories (access control, encryption, logging, key management, risk management, and incident management) covering the main aspects of information security. The overall privacy indicator of the system was calculated using a weighted sum, where the weighting coefficients were refined depending on the criticality of each parameter. To unify the scales and ensure correct further analysis, normalisation methods (minimax and Z-normalisation) were applied, thanks to which the obtained parameter values can be compared and effectively integrated into the general model. The proposed method used differential privacy to protect source data and enhance privacy, which was achieved by adding random noise with a normal distribution. This step complicated the process of restoring the original indicators and minimised the risk of identifying specific records, while maintaining the accuracy of aggregate statistical estimates. The developed approach consisted of several sequential stages: from initial data categorisation and normalisation to the

Suggested Citation:
Prokopovych-Tkachenko, D., Rybalchenko, L., Zvieriev, V., Khrushkov, B., & Bushkov, V. (2025). Integrated assessment of system privacy: Formalisation, normalisation and differential privacy. *Information Technologies and Computer Engineering*, 22(3), 125-135. doi: 10.31649/vitce/3.2025.125

*Corresponding author

implementation of differential privacy and data analysis in a neural network. An important advantage was the ability to integrate various aspects of data protection into a single coherent system. This multidimensional concept promoted flexibility and allowed the solution to be quickly adapted to updated requirements or new threats. The presented model is particularly relevant in areas where sensitive data is processed: healthcare, banking and finance, as well as public administration and information security. The proposed approach lays the foundation for the development and scaling of secure and transparent systems that meet modern privacy standards

Keywords: information protection; secure neural networks; vector data normalisation; access to information; security assessment parameters; statistical noise; adaptive machine learning

Introduction

In the modern era of rapid digital transformation, the volume of data generated, transmitted and processed by information systems is growing so quickly that traditional protection approaches can no longer reliably meet confidentiality and privacy requirements. This issue is particularly pressing in areas where sensitive data are processed, such as healthcare, the financial sector, public administration and systems dealing with the security of critical information. In such contexts, it is crucial to choose comprehensive protection methods and integrate them into a unified system, taking into account the diversity of confidentiality parameters that affect the overall security level, as well as the fast-changing nature of threats and new regulatory requirements. Given the dynamics of technologies and threats, there is a need for methods capable not only of analysing current security conditions but also of adapting flexibly to the emergence of new challenges.

Among the key trends in ensuring confidentiality is the combination of deep learning methods with neuro-symbolic artificial intelligence. Thus, A. Piplai *et al*. (2023) described an innovative approach that integrates knowledge graphs and deep learning to enhance the interpretability of models in the field of cybersecurity. The integration of neural networks represents knowledge of the relevant subject area and allows artificial intelligence (AI) systems to reason, learn, and generalise in a way that is understandable to experts.

Among Ukrainian sources, O.M. Gumen & K.O. Rachek (2023) are noteworthy for their use of machine learning to predict space weather while ensuring privacy. This article discusses models for predicting the Dst index. One of the models uses a precision of 83.47%. Another model, Dst Transformer (DSTT), is designed for short-term forecasting and uses Bayesian inference. The DSTT model shows high accuracy and takes into account two types of uncertainties in the data. I. Grinko *et al*. (2023) presented an overview of quantum convolutional networks for interdisciplinary use, particularly in socio-economic systems. Modelling and forecasting complex natural processes has demonstrated their effectiveness in studying complex molecular structures. It has been established that quantum convolutional neural networks can provide more accurate and faster results compared to conventional data processing methods. The work of N. Zaplatynskyi *et al*. (2024) emphasised that the growth in data volumes and the increasing complexity of information flows require comprehensive approaches to their processing and protection, including the use of AI, and that confidentiality must be integrated at the system architecture level.

In response to growing privacy requirements in distributed data processing environments, integrated information security assessment is actively used in training. R. Shokri *et al*. (2017) demonstrated the danger of training models without privacy mechanisms. The CYBRIA development, presented by P. Thantharate & T. Anurag (2023), allows models to be trained without sharing raw data, which significantly reduces the risk of information leakage. This article describes how eco-symbolic AI can be useful in the fields of cybersecurity and privacy – two of the most demanding areas where AI must be understandable and at the same time highly accurate in complex conditions. A similar approach is taken in the study by S. Sav *et al*. (2023), which demonstrated the effectiveness of federated recurrent networks with privacy in mind. The researchers pay particular attention to differential privacy as a means of protecting personal data. H. Lee *et al.* (2023) proved the effectiveness of adding Gaussian noise in industrial data processing tasks. The MNP method has shown significant potential for making production systems both intelligent and secure, eliminating the risk of data leakage while maintaining the quality of AI models.

The feasibility of a comprehensive study of cybersecurity issues was presented in the work of O. Chubukova *et al*. (2020), which applies machine learning algorithms and risk identification features that occur in the banking sector, namely through the use of data science to detect fraud and model risks for investment institutions. The analysis of problem areas was investigated by V. Ivanichenko *et al.* (2021). The work uses machine learning in cybersecurity to implement important issues of creating a self-learning model for reliable protection in information security decision-making. O. Semenenko *et a*l. (2024) proved that integrated computer technologies increase the level of cybersecurity in the defence sector by ensuring the detection of and response to cyber threats. Regarding the prevention of security breaches, M.A. Fathullah *et al*. (2023) proposed cloud computing mechanisms using IT projects to control and prevent risks, threats, vulnerabilities, probabilities, consequences, and control procedures, which are

classified into separate risk classes for further management decision-making.

Thus, the issue of risk reduction and data privacy remains relevant, attracting increased interest from scientists and software developers. Consequently, there was still a need to develop a methodology based on a mathematical model that involves preliminary data normalisation using a multi-layer neural network for classification. The aim of the current study was to develop a formalised model for the integrated assessment of information system confidentiality, combining a mathematical representation of security parameters, multi-level normalisation, differential privacy mechanisms and the use of neural networks to ensure the protection of sensitive data in the context of modern cyber threats.

Materials and Methods

The study was conducted using a general methodology for building privacy assessment systems adapted to modern data protection challenges. The methodology included the sequential implementation of four stages. The first stage involved categorising parameters that reflect the main areas of confidentiality assurance: access control, encryption, logging, key management, risk and incident management, etc. This approach made it possible to systematically structure the characteristics of the system and identify critical areas. The second stage involved assessing the weighting coefficients of the parameters, taking into account industry criticality, the probability of threats being realised, and the consequences of their impact. An approach consistent with risk management practices in cloud environments was applied, as well as basic approaches to weighted analysis. In the third stage, data normalisation was performed using minimax and Z-transformation, which allowed the parameters to be standardised for further calculation and analysis. This ensured data compatibility for use in intelligent models. The fourth stage involved the implementation of differential privacy mechanisms. To do this, Gaussian or Laplace noise components were added to the normalised data, in accordance with current personal data protection practices.

In the final stage, a multilayer perceptron neural network (MLP) was used as the base model for classifying the confidentiality level of systems. This type of neural network is a classic form of a deep feedforward neural network, which consists of:

- an input layer that accepts vectorised privacy parameters;
- one or more hidden layers that implement nonlinear transformations;
- an output layer that forms the final assessment of the privacy level or classification (e.g., “low,” “medium,” “high” level).

The reasons for choosing MLP were: adaptability to different types of data after normalisation; compatibility with differential privacy mechanisms (especially when using DP-SGD); high accuracy in classification tasks under noise conditions. The use of this approach made it possible to form a consistent privacy assessment system with adaptive properties and compliance with privacy requirements in the fields of healthcare, finance, information risk management, and recommendation systems.

Mathematical representation and formalisation of confidentiality parameters

The differential concept of system confidentiality is defined by a set of parameters, which can be denoted as:

$$P = \{p_1, p_2, p_3, \ldots, p_n\} \tag{1}$$

Then these parameters can be represented as a vector:

$$P = [p_1, p_2, p_3, \ldots, p_n]. \tag{2}$$

Each parameter p_i may take a numerical value within a defined interval (for example, [0; 1]), reflecting the degree of its implementation or importance.

Parameter categories. For convenience, the entire set can be divided into subsets (categories), for example: P_1 = {Access Control}, P_2 = {Encryption}, P_3 = {Logging}. Then the overall set of parameters is:

$$P = P_1 \cup P_2 \cup P_3 \cup \cdots \cup P_m. \tag{3}$$

The overall privacy assessment function $F(p)$ evaluates the level of system privacy based on the parameter vector p. The overall assessment can then be expressed as a weighted sum of all parameters:

$$F(p) = w_1 \cdot p_1 + w_2 \cdot p_2 + \ldots + w_n \cdot p_n, \tag{4}$$

where w_i – the weighting coefficient reflecting the importance of the corresponding parameter p_i.

Data normalisation for training. Before input data are fed into the neural network, all parameters p_i are normalised. This means that each parameter value is brought to a common scale, for example to the interval [0; 1], to ensure correct processing in the model and to improve its convergence during training:

$$p_i^{norm} = \frac{p_i - p_i^{min}}{p_i^{max} - p_i^{min}}. \tag{5}$$

This makes the values comparable and improves convergence during training.

Differential privacy (optional). To protect privacy during training, noise may be added:

$$p_i^{dp} = p_i + \mathcal{N}(0, \sigma^2), \tag{6}$$

where $N(0, \sigma^2)$ – is noise distributed according to a normal (Gaussian) distribution with mean 0 and variance σ^2.

The applied model enabled the prediction of protection and security levels, as well as timely anomaly detection using neural networks with high data accuracy. The privacy-assessment model was implemented taking into

account a formalised parameter structure, multi-level normalisation, differential privacy mechanisms and the integration of a neural network for processing protected data. Based on predefined weighting coefficients, structured by categories, and using an adaptive multilayer perceptron architecture, an experimental evaluation of the model's effectiveness was carried out. During the analysis, particular attention was paid to prediction accuracy, sensitivity to normalisation parameters, and the impact of the noise level introduced according to differential privacy requirements. The results made it possible to assess the practical feasibility of applying the developed approach in systems operating under conditions involving the processing of sensitive or personalised information.

Results and Discussion

Summary of privacy assessment by category

Effective privacy assessment requires not only qualitative analysis of security parameters, but also a formalised approach to their structuring, weighting and processing. Thus, a structured approach to organising parameters was used, which ensured flexibility, scalability and logical integrity of the analysis process. The assessment parameters were grouped into categories (labelled as $X^{(k)}$, each of which corresponds to a separate aspect of privacy – for example, "Access Control" and so on. This division allows for a systematic coverage of all important security areas, simplifies comparisons between systems with different architectures and security policies, and creates a basis for a differentiated approach to assessment, where certain categories may be given greater weight depending on the context of application. For example, a set of parameters:

$$X = \{X^{(1)}, X^{(2)}, \ldots, X^{(C)}\}, \tag{7}$$

where C – number of main categories ("Access Control", "Encryption", "Logging", etc.)

Each category $X^{(k)}$ is itself a subset (or vector) of parameters that describe specific characteristics or security settings within that category, providing detail down to the level of individual components:

$$X^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \ldots, x_{n_k}^{(k)}\right), \tag{8}$$

where n_k – the number of parameters in category k.

This allows to introduce a separate group of weight coefficients for each category:

$$W^{(k)} = \left(w_1^{(k)}, w_2^{(k)}, \ldots, w_{n_k}^{(k)}\right), \tag{9}$$

as well as one "global" weight coefficient α_k, which reflects the importance of the entire category:

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_C). \tag{10}$$

Then, at the simplest level (linear model), it can be written as:

$$f(X) = \sum_{k=1}^{C} \alpha_k \left(\sum_{i=1}^{n_k} w_i^{(k)} x_i^{(k)}\right). \tag{11}$$

This allows to take into account differences in anomalies that are possible between different groups of parameters that are important for privacy and creating reliable security, as well as to adapt the model to specific system requirements or regulatory restrictions (for example, in the banking sector, encryption may be more important than logging). This approach helps to balance detail and generalisation at the integrated assessment level. Thus, the use of a categorically structured parameter model in combination with a weighting system and multi-level normalisation provides both flexibility and formal justification for the privacy assessment process.

Normalisation and standardisation at several levels

The unification of input data makes it possible to increase the accuracy and stability of calculations, as well as to ensure the comparability of parameters coming from different sources and belonging to different privacy categories. At this stage, the parameters for further research were normalised and standardised, which was implemented at several interrelated levels. This made it possible to perform calculations to identify favourable directions for ensuring privacy.

Instead of simple standardisation $x_i' = \frac{x_i - \mu_i}{\sigma_i}$ an extended approach can be used.

1. Normalisation within a category:

$$x_i^{(k)} \mapsto \tilde{x}_i^{(k)} = \frac{x_i^{(k)} - \mu_i^{(k)}}{\sigma_i^{(k)}}, \tag{12}$$

where $\mu_i^{(k)}, \sigma_i^{(k)}$ are calculated only for category k.

2. Normalisation between categories: if there are categories in which the parameter values have different scales, additional global correction or scaling can be performed.

3. Limiting values (e.g., using a sigmoid or other nonlinear function):

$$\hat{x}_i^{(k)} = \sigma\left(\tilde{x}_i^{(k)}\right) = \frac{1}{1 + e^{-\tilde{x}_i^{(k)}}}. \tag{13}$$

This ensures that all reduced values lie in the interval [0,1].

Thus, the complicated version of the input data may be as follows:

$$\hat{X}^{(k)} = \left(\hat{x}_1^{(k)}, \hat{x}_2^{(k)}, \ldots, \hat{x}_{n_k}^{(k)}\right) \tag{14}$$

and instead of $x_i^{(k)}$ is now used in the formula $\hat{x}_1^{(k)}$.

After bringing the input parameters to a unified scale, it is necessary to take into account the relationships between them, which can significantly affect the accuracy of the privacy assessment. Simple weighting does not always reflect the real complexity of the dependencies between individual security characteristics, especially in conditions of high data density. That is why the next stage of the model was to expand the computational scheme to take into account internal and inter-category correlations.

Taking into account the correlation between parameters

Cross-correlation in the context of categories means researching and identifying the correlation between different categories. That is, determining how closely different categories are related to each other. The correlation can be either negative or positive. This makes it possible to identify the relationship between factors that influence indicators that point to dangerous manifestations and threats. If the parameters within a category or between categories influence each other, quadratic or interaction terms can be added. For example, in each category k, instead of the sum $\sum_{i=1}^{n_k} w_i^{(k)} \hat{x}_i^{(k)}$ a generalised expression can be applied:

$$\sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \beta_{ij}^{(k)} \hat{x}_i^{(k)} \hat{x}_j^{(k)}, \quad (15)$$

where $\beta^{(k)}$ – a matrix of parameters (weights) that takes into account: diagonal elements $\beta_{ij}^{(k)}$ correspond to the "strength" of the influence of a single parameter; non-diagonal elements $\beta_{ij}^{(k)}$ reflect the mutual influence of a pair of parameters i, j.

Then the model in category k will look like this:

$$g_k\left(\hat{X}^{(k)}\right) = \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \beta_{ij}^{(k)} \hat{x}_i^{(k)} \hat{x}_j^{(k)}. \quad (16)$$

And the overall estimate:

$$f(X) = \sum_{k=1}^{C} \alpha_k g_k\left(\hat{X}^{(k)}\right). \quad (17)$$

If even greater accuracy is required, cross-correlations between categories can be used:

$$\sum_{k=1}^{C} \sum_{m=1}^{C} \sum_{i=1}^{n_k} \sum_{j=1}^{n_m} \gamma_{k,m}^{(i,j)} \hat{x}_i^{(k)} \hat{x}_j^{(m)}. \quad (18)$$

This significantly increases the number of parameters to be adjusted. This means that a large data set can be used for research, which improves the quality of the research itself and the final results. Taking into account the correlation between parameters allows the model to more accurately reflect the relationships within and between categories, which is important for a comprehensive assessment of privacy. Extending the basic linear model with quadratic and cross-categorical terms allows to identify both the individual and combined effects of parameters on the overall level of security. This approach creates conditions for flexible adaptation of the model to complex information system structures.

Regularisation and restrictions on weight coefficients

For further research, it was advisable to introduce restrictions on weight coefficients. This made it possible to select from the general population those indicators that have a significant impact on the factors of privacy and to reduce the influence of insignificant factors. To avoid an "explosion" of parameters or an overestimation of the impact of specific characteristics, regularisation is often used:

1. L2 regularisation (ridge regression): adds the sum of the squares of the weights to the loss function. For example, if Θ – is the set of all $\alpha_k, w_i^{(k)}, \beta_{ij}^{(k)}$ or, $\gamma_{k,m}^{(i,j)}$ then:

$$\Omega(\Theta) = \lambda \sum_{\theta \in \Theta} \theta^2, \quad (19)$$

where λ – a hyperparameter that determines the "strength" of regularisation.

2. L1 regularisation (lasso regression): inclines some weights towards zero, which effectively cuts off insignificant parameters:

$$\Omega(\Theta) = \lambda \sum_{\theta \in \Theta} |\theta|. \quad (20)$$

This helps to obtain results for further research from those factors that have a significant impact on the indicators.

3. Weight sum constraint: it is possible to require that $\sum_{i=1}^{n_k} w_i^{(k)} = 1$ (or a similar constraint for $\beta_{ij}^{(k)}$), to ensure a certain "normality" of influence.

4. Restricting the signs of weights (e.g., $\alpha_k \geqslant 0$).

When training or calibrating a model, the total loss function (e.g., L) may contain both deviations from the desired "correct" values and regularisation:

$$L(\Theta) = Loss(\Theta) + \Omega(\Theta). \quad (21)$$

The study conducted on calculating restrictions on weight coefficients makes it possible to reduce the number of indicators selected from the total amount of data and significantly influence those factors that are closely related to the factors for assessing the confidentiality of information systems. Normalisation and limitation of weight coefficient signs ensure the interpretability and consistency of results. Confidentiality parameters, combined with weight coefficients, form a differentiated model suitable for preparing data for neural network training.

An extended approach to differential privacy

Adding noise to parameters. To ensure formal privacy guarantees during data processing and analysis, mechanisms must be implemented to reduce the risk of confidential information leaks. One of the key approaches in this area is to add random noise to input parameters or calculation results, which makes it more difficult to identify individual records. This method allows the principles of differential privacy to be implemented, ensuring a balance between model accuracy and data protection. Earlier, a simple scheme was mentioned for adding Gaussian noise to normalised $\tilde{x}_i$. However, in practice, differential privacy often uses Laplace, Gaussian, and functional mechanisms.

Laplace mechanism:

$$\tilde{x}_i = \hat{x}_i + Laplace(0, b), \quad (22)$$

where Laplace $(0, b)$ – noise from the Laplace distribution, determined by parameter b.

Gaussian mechanism:

$$\tilde{x}_i = \hat{x}_i + \mathcal{N}(0, \sigma^2), \quad (23)$$

where σ selected to ensure (ε, δ) – differential privacy.

Functional mechanism: if it is not the vector X, itself that is calculated, but the result of some function $f(X)$, noise is added directly to the output of the function:

$$f(X) \mapsto f(X) + \eta, \quad (24)$$

where η selected from the desired distribution, depending on the sensitivity of f.

Adding noise at the gradient stage (DP-SGD) – an effective method for ensuring differential privacy during neural network training. The idea is to modify the standard stochastic gradient descent (SGD) algorithm and add random Gaussian noise. If weights Θ (e.g., α_k, $\beta_{ij}^{(k)}$ etc.) are trained using stochastic gradient descent, differential privacy can be implemented through the "Clip Noise" mechanism by introducing two key mechanisms: gradient $\nabla L(\Theta)$ clipping (each gradient reduced to a limited norm, for example, $\|\nabla L(\Theta)\| \leqslant k$) and adding random Gaussian noise:

$$\nabla L(\Theta) \mapsto \nabla L(\Theta) + \mathcal{N}(0, \sigma^2 k^2). \quad (25)$$

This noise prevents the accurate reconstruction of individual data contributions to the gradient, providing a formal guarantee of differential privacy. The combination of these two steps allows to control the balance between training quality (model accuracy) and privacy protection. Increasing the parameter σ improves protection but may reduce model performance, requiring careful tuning. Thus, the approach to differential privacy can be more complex than simply "adding noise to the parameters".

Example of a generalised formula for privacy assessment. Taking all of the above into account, the sequence of the "extended" privacy assessment takes the following form:

1. Normalisation (preliminary step at the category level):

$$\hat{x}_i^{(k)} = NonlinearNorm\left(x_i^{(k)}\right), \quad (26)$$

where *NonlinearNorm* – a composite scaling procedure.

2. Interactions and weights (within category):

$$g_k\left(\hat{X}^{(k)}\right) = \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \beta_{ij}^{(k)} \hat{x}_i^{(k)} \hat{x}_j^{(k)}. \quad (27)$$

3. Global weight of category α_k.

4. Intercategory term (optional):

$$h(X) = \sum_{k=1}^{C} \sum_{m=k+1}^{C} \sum_{i=1}^{n_k} \sum_{j=1}^{n_m} \gamma_{k,m}^{(i,j)} \hat{x}_i^{(k)} \hat{x}_j^{(m)}. \quad (28)$$

5. Conclusion:

$$f(X) = \underbrace{\sum_{k=1}^{C} \alpha_k g_k\left(\hat{X}^{(k)}\right)}_{\text{intra-category interactions}} + \underbrace{h(X)}_{\text{inter-category interactions}} + \varepsilon_{noise}, \quad (29)$$

where ε_{noise} – random noise that can be added: directly to $f(X)$ (Laplace/Gaussian mechanism); during the training of β, α, γ (via DP-SGD).

6. Regularisation (etc.) added to the loss function during training or calibration $\alpha_k, \beta_{ij}^{(k)}, \gamma_{k,m}^{(i,j)}$.

Thus, unlike the basic linear combination with normalisation, the "extended" scheme contains:

- grouping of parameters into categories with their own coefficients, plus global coefficients for the entire block;
- non-linear transformations and multi-level normalisation (internal and global);
- accounting for cross-influences: through additional matrices β, γ;
- regularisation to prevent overfitting and inadequately large weights;
- differential privacy (through noise in the data, in the output function, or in the gradient during training).

All this increases the complexity of calculations and the number of parameters, but allows for more flexible modelling of dependencies between security/privacy factors and, if necessary, protects the privacy of final data or intermediate results. Below is an example of a simple block diagram (Fig. 1) that shows the main categories of privacy parameters (access control, encryption, logging) and how they are combined into a general evaluation function. The diagram demonstrates the process of processing confidential parameters aimed at ensuring data privacy and improving data security. The visualisation shows a multi-stage approach to information processing: from data categorisation to normalisation, ensuring differential privacy and further analysis using a neural network. This approach allows different aspects of data protection to be integrated into a single structure that takes into account the requirements of modern confidential information processing systems.

The described block diagram is an important tool for developers and analysts, as it helps visualise and understand complex processes involved in handling confidential information. It allows potential vulnerabilities and weak points within the system to be easily identified, as well as helping to optimise data-protection strategies. The block diagram serves not only as an internal tool during the development process but also as a means of communication with clients and other stakeholders, helping to understand the importance of privacy and data security in the modern digital environment.

All input data is divided into categories that reflect key aspects of information security. Access control ensures that data access rights are verified. Encryption implements mechanisms to protect information from unauthorised access. Logging is responsible for collecting and storing data about events in the system. Key management includes operations for storing, processing, and rotating encryption keys. Risk management assesses and monitors potential threats. Incident management focuses on detecting and responding to security incidents. After collecting data from

different categories, it is combined into a single metric using a weighted sum. Weighting coefficients reflect the importance of each parameter, and the parameter values are combined in formula (9). The resulting metric is normalised to bring it to a single scale. Min-max normalisation is used, which brings the values to a range from minimum to maximum, or Z-normalisation, which is based on the mean and standard deviation. To ensure privacy, random noise is added to the normalised values according to the normal distribution law. This makes it difficult to recover the original data, protecting privacy and compliance with differential privacy requirements. In the final stage, the processed data is fed into a neural network, which is used for prediction, classification, or other data analysis tasks.

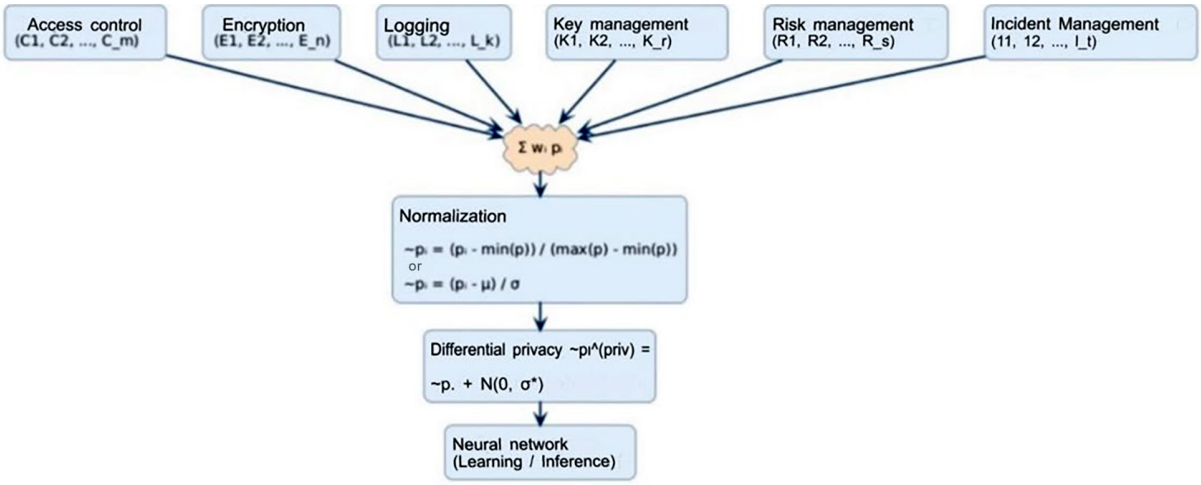


Figure 1. Block diagram of the privacy-parameter category
Source: constructed based on data from the authors

This scheme reflects an integrated approach to the processing of confidential data that can be applied in systems focused on the protection of personal information, including in the fields of healthcare, finance, and information security. The described privacy parameters, together with weighting coefficients, allow to construct a differential concept of privacy. Such formalisation will be useful when preparing and normalising data for further training of neural networks or other machine learning methods.

The described scheme is an important step towards ensuring data privacy and security in the modern world, where information is a critically valuable resource. The use of such data processing methods not only protects personal information from unauthorised access but also ensures compliance with legal requirements and ethical standards. The developed methodology can be adapted and used in various fields where confidential data processing is required, contributing to an increase in user confidence in information processing systems and ensuring the stability and security of their operation.

Differential privacy is one of the most effective approaches to data privacy protection, particularly in areas where sensitive information is processed. Its basic principle is to add random noise to data or calculation results, making it impossible to recover the original values. Formally, differential privacy is achieved by adding noise that has a normal or Laplace distribution. For example, for a normalised parameter x_i, the following is applied:

$$x_i' = x_i + \mathcal{N}(0, \sigma^2), \tag{30}$$

where $N(0, \sigma^2)$ – random noise with normal distribution, parameter σ controls the level of data blurring.

In recommendation systems for search and streaming services, algorithms analyse user preferences. Based on formula (22), adding noise to the output data is provided by the following formula:

$$r' = r + Lap(\lambda), \tag{31}$$

where $Lap(\lambda)$ – Laplace noise with parameter λ, which determines the level of privacy.

In the financial sector, for secure analysis of transaction data, the total number of transactions can be calculated using:

$$T' = T + \mathcal{N}(0, \sigma_T^2), \tag{32}$$

where T – the actual number of transactions, and σ_T^2 controls the level of differential privacy.

To reduce the impact of noise on the accuracy of the analysis, adaptive mechanisms can be used to adjust the privacy parameters. For example, the gradient noise mechanism (*DP-SGD*):

$$\tilde{g}_i = clip(g_i, C) + \mathcal{N}\left(0, \sigma_g^2\right), \tag{33}$$

where g_i – the gradient of the loss function; C – the threshold value (clipping); $N(0, \sigma^{g2})$ – additional Gaussian noise.

Thus, the integration of differential privacy allows for secure information analysis algorithms. It is important

to note that differential privacy does not guarantee the absolute impossibility of identifying an individual, but it makes this process extremely difficult and minimises the risk of confidential information leakage. It is for this reason that differential privacy is a powerful tool for protecting personal data in the modern digital world, where the processing of large amounts of information is the norm (Dwork & Roth, 2014).

Differential privacy is one of the most effective approaches to protecting data confidentiality, especially in areas where sensitive information is processed. Its basic principle is to add random noise to the input data or to the results of calculations, making it impossible to accurately recover the original values. This allows aggregated data to be used for analysis while maintaining the anonymity of individual records. One of the most common methods of ensuring differential privacy is the use of random noise, which can have a normal or Laplace distribution. For example, when analysing normalised parameters, a random component distributed according to a specific law is added to their values. The degree of data blurring is determined by the corresponding distribution parameters, which control the level of confidentiality.

In recommendation systems used in search and streaming services, algorithms analyse user preferences to improve personalised recommendations. Adding random noise to records of user interactions with content ensures privacy with little impact on system performance. Thus, the confidentiality of personal preferences is preserved without significantly reducing the effectiveness of the algorithm. In the financial sector, differential privacy is used to analyse transaction data, allowing banks to identify suspicious transactions without revealing information about specific customers. This is achieved by modifying aggregate metrics, such as the total number of transactions over a given period, by adding random noise. Since adding noise can affect the accuracy of the analysis results, it is important to choose adaptive mechanisms that allow the level of privacy to be adjusted according to specific needs. For example, when using neural networks, methods of regulating gradient weight updates can be applied by adding random noise during the model training stage. This reduces the risk of recovering the original data from intermediate results, with little impact on the performance of the algorithm.

This study uses a multilayer neural network as a baseline model for classifying the privacy level of systems, indicating its effectiveness for data protection. These results are also confirmed by studies that used other approaches. In particular, S. Tyshchenko & E. Kuznetsov (2024) described the use of neural networks in image classification. The authors solved the problem based on the task of entering an image into a neural network and assigning any label to the image. A time-efficient dataset was used to build the training model, which depended on the size and quality of the dataset. A. Rutkas & V. Shtanko (2024) raised a philosophical question about the importance of using artificial neural networks for interaction between humans and artificial systems. This idea has been technically developed in the current study, as the integration of differential privacy not only provides a technical level of security but also increases user confidence in automated data analysis systems.

In some publications, the authors focused on the applied economic use of artificial neural networks: N. Savka *et al*. (2020) analyse the forecasting of business activity using radial basis networks. The performance indicators of an enterprise depend on the specifics of its marketing policy, which is especially important for enterprises involved in product sales. Most existing methods for modelling enterprise activity are based on statistical and mathematical methods. Similarly, the proposed current methodology has demonstrated flexibility, allowing the model to be adapted to the specifics of specific domains – from finance to healthcare. H. Liavynets *et al*. (2024) investigated the application of neural network models in the hotel and restaurant industry for processing and analysing large amounts of data, which makes it possible to forecast information for strategic management decisions in the hotel and restaurant business.

The work of Y. Terpilovskyi (2024) explores bioinformatics and the representation of *k*-mer DNA data. The first method used by the author employs a vector of binary features, where each possible *k*-mer corresponds to a binary feature, resulting in high-dimensional and sparse feature vectors. The second method was based on the Conway-Bromage-Lyndon (CBL) structure, which introduces a compressed and dynamic representation of k-mers. In the proposed study, the problem of sparsity and multidimensionality is solved by normalising parameters and introducing noise mechanisms, which allows confidentiality to be maintained without losing informativeness.

In the study by A. Volokyta & M. Melenchukov (2024), neural networks are used to detect attacks in distributed systems. In this context, the proposed methodology demonstrates potential in the field of cyber security, especially given its scalability and applicability in public administration systems and financial infrastructure. In the current study, a neural network is used to enhance data protection confidentiality. Equally revealing is the analysis by M.S. Ahsan & A.-S.K. Pathan (2025), who draw attention to the security issues of the Internet of Things. One of the key issues is the identification of potential vulnerabilities and access control, which determines the overall security of Internet of Things systems. These tasks are also solved using the developed approach thanks to a multi-level risk assessment structure. A distinctive feature of the current study was the use of a protection prediction model for the timely detection of anomalies, which uses neural networks with high data accuracy.

Thus, the study highlighted the growing role of comprehensive approaches to privacy assessment that combine mathematical formalisation, adaptive data processing methods, and differential privacy. Analysis of the literature confirmed that effective protection of information systems requires interdisciplinary integration of technical, organisational, and ethical solutions. In summary, the results

demonstrate that combining neural network models with differential privacy mechanisms is a promising direction for creating robust and reliable data protection systems.

Conclusions

This study developed a formalised approach to the integrated assessment of information system privacy, taking into account the multi-component structure of risks and modern requirements for data privacy. The proposed methodology is based on the mathematical representation of parameters in the form of vectors, subsequent data normalisation, the application of weighting coefficients, and the implementation of differential privacy. Within the scope of the study, confidentiality parameters were structured by category, multi-level normalisation was performed, noise addition mechanisms were implemented at the processing and training stages, and a multi-layer neural network was used for classification. The results confirm the effectiveness of the developed model in maintaining a balance between the accuracy of analytical forecasts and the level of protection of confidential information.

The method provides flexibility in configuring the structure of weight coefficients, allows taking into account both the criticality of individual parameters and their categorical significance, and also allows scaling the system for different application domains. The inclusion of differential privacy mechanisms, in particular the addition of noise (Laplace, Gaussian, functional mechanism) and the use of DP-SGD during training, increases the level of privacy and makes the approach relevant for modern automated information protection systems. The developed block diagram, which reflects the process of categorisation, normalisation, differential privacy and further analysis of parameters, demonstrates the practical applicability of the proposed approach for developers and analysts. It facilitates the identification of potential vulnerabilities, the optimisation of protection strategies and communication with all interested parties.

The approach proposed in the study is relevant for use in healthcare, the financial sector, public administration systems, recommendation systems, and other areas where the secure processing of personalised data is important. The results conceptualise the possibility of creating unified privacy assessment systems based on quantitative criteria and modern artificial intelligence algorithms. This highlights the practical significance of the approach for building reliable, transparent, and ethical information systems that meet digital security and regulatory compliance requirements. This approach allows maintaining a balance between data security and the ability to perform financial analysis.

Thus, the integration of differential privacy into data analysis systems ensures their security and compliance with modern confidentiality requirements. The use of methods for adding random noise, adaptive privacy level control, and algorithm parameter adjustment allows for the creation of effective information processing mechanisms without the risk of sensitive data disclosure. Promising areas for further research include expanding the model to take into account context-oriented risk parameters, integration with behavioural analysis systems, and optimisation of adaptive privacy level control mechanisms depending on the user profile and the type of data being processed. Modifying aggregate indicators by adding random noise allows maintaining a balance between data security and the ability to perform analysis.

Acknowledgements

None.

Funding

The study was not funded.

Conflict of Interest

None.

References

[1] Ahsan, M.S., & Pathan, A.-S.K. (2025). A comprehensive survey on the requirements, applications, and future challenges for access control models in IoT: The state of the art. *IoT*, 6(1), article number 9. doi: 10.3390/iot6010009.

[2] Chubukova, O., Ponomarenko, I., & Domantovych, O. (2020). Using data science to risk assessment. *Market Infrastructure*, 47, 129-132. doi: 10.32843/infrastruct47-24.

[3] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4), 211-407. doi: 10.1561/0400000042.

[4] Fathullah, M.A., Subbarao, A., & Muthaiyah, S. (2023). A systematic review: Risk management of cloud computing projects in healthcare. *International Journal of Management, Finance and Accounting*, 4(2), 83-115. doi: 10.33093/ijomfa.2023.4.2.5.

[5] Grinko, I., Skrypnyk, T., & Barmak, O. (2023). Quantum convolutional neural networks: Features of implementation in technical, natural and socio-economic systems. *Herald of Khmelnytskyi National University. Technical Sciences*, 323(4), 87-94. doi: 10.31891/2307-5732-2023-323-4-87-94.

[6] Gumen, O.M., & Rachek, K.O. (2023). Neural networks and machine learning in data processing for space weather forecasting. *Applied Questions of Mathematical Modeling*, 6(2), 19-23. doi: 10.32782/mathematical-modelling/2023-6-2-2.

[7] Ivanichenko, V., Sablina, M., & Kravchuk, K. (2021). Use of machine learning in cyber security. *Cybersecurity: Education, Science, Technology*, 4(12), 132-142. doi: 10.28925/2663-4023.2021.12.132142.

[8] Lee, H., Finke, D.C., & Yang, H. (2023). Privacy-preserving neural networks for smart manufacturing. *Journal of Computing and Information Science in Engineering*, 24(7), article number 071002. doi: 10.1115/1.4063728.
[9] Liavynets, H., Liulka, O., & Tkachuk, Y. (2024). Shallow artificial neural networks in management hotel and restaurant business. *Economy and Society*, 68. doi: 10.32782/2524-0072/2024-68-46.
[10] Piplai, A., Kotal, A., Mohseni, S., Gaur, M., Mittal, S., & Joshi, A. (2023). Knowledge-enhanced neurosymbolic artificial intelligence for cybersecurity and privacy. *IEEE Internet Computing*, 27(5), 43-48. doi: 10.1109/MIC.2023.3299435.
[11] Rutkas, A., & Shtanko, V. (2024). Artificial neural networks: A tool or a partner of the human mind. *Grail of Science*, 47, 652-659. doi: 10.36074/grail-of-science.20.12.2024.099.
[12] Sav, S., Diaa, A., Pyrgelis, A., Bossuat, J.-P., & Hubaux, J.-P. (2023). Privacy-preserving federated recurrent neural networks. *Proceedings on Privacy Enhancing Technologies*, 2023(4), 500-521. doi: 10.56553/popets-2023-0122.
[13] Savka, N., Vasylkiv, N., Dubchak, L., & Mudryk, I. (2020). Radial-basis neural networks for enterprises activity prediction. *European Science*, 3(sge17-03), 42-48. doi: 10.30890/2709-2313.2023-17-03-012.
[14] Semenenko, O., Kirsanov, S., Movchan, A., Ihnatiev, M., & Dobrovolskyi, U. (2024). Impact of computer-integrated technologies on cybersecurity in the defence sector. *Machinery & Energetics*, 15(2), 118-129. doi: 10.31548/machinery/2.2024.118.
[15] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *IEEE symposium on security and privacy* (pp. 3-18). San Jose: IEEE. doi: 10.1109/SP.2017.41.
[16] Terpilovskyi, Y. (2024). Comparison of DNA k-mer data representations for classification via neural networks. *International Scientific Technical Journal "Problems of Control and Informatics"*, 69(6), 61-69. doi: 10.34229/1028-0979-2024-6-5.
[17] Thantharate, P., & Anurag, T. (2023). CYBRIA – Pioneering federated learning for privacy aware cybersecurity. In *IEEE 20th international conference on smart communities: Improving quality of life using AI, robotics and IoT (HONET)* (pp. 56-61). Boca Raton: IEEE. doi: 10.1109/honet59747.2023.10374608.
[18] Tyshchenko, S., & Kuznetsov, E. (2024). Neural networks for the problem of image classification. *Science and Technology Today*, 3(31). doi: 10.52058/2786-6025-2024-3(31)-705-718.
[19] Volokyta, A., & Melenchukov, M. (2024). Neural networks in detecting attacks on distributed systems. *Technical Sciences and Technologies*, 1(35), 135-145. doi: 10.25140/2411-5363-2024-1(35)-135-145.
[20] Zaplatynskyi, N., Lub, P., & Zaporozhtsev, S. (2024). Improving cybersecurity with artificial intelligence. *Bulletin of Cherkasy State Technological University*, 29(4), 53-61. doi: 10.62660/bcstu/4.2024.53.

Інтегрована оцінка конфіденційності систем: формалізація, нормалізація та диференційна приватність

Дмитро Прокопович-Ткаченко

Кандидат технічних наук, доцент
Університет митної справи та фінансів
49000, вул. Володимира Вернадського, 2/4, м. Дніпро, Україна
https://orcid.org/0000-0002-6590-3898

Людмила Рибальченко

Кандидат економічних наук, доцент
Університет митної справи та фінансів
49000, вул. Володимира Вернадського, 2/4, м. Дніпро, Україна
https://orcid.org/0000-0003-0413-8296

Володимир Звєрєв

Кандидат технічних наук, доцент
Державний торговельно-економічний університет
02156, вул. Кіото, 19, м. Київ, Україна
https://orcid.org/0000-0002-0907-0705

Борис Хрушков

Аспірант
Університет митної справи та фінансів
49000, вул. Володимира Вернадського, 2/4, м. Дніпро, Україна
https://orcid.org/0009-0002-3978-5012

Бушков Валерій

Аспірант
Державний торговельно-економічний університет
02156, вул. Кіото, 19, м. Київ, Україна
https://orcid.org/0009-0005-5097-2689

Анотація. Вимоги щодо конфіденційності та приватності даних дедалі більше зростають. Метою роботи було розробити формалізований підхід до оцінювання конфіденційності інформаційних систем, що базується на векторному поданні множини параметрів. У запропонованому підході кожен параметр має числове значення у визначеному інтервалі, яке відображає ступінь його реалізації або важливості. Для зручності та структурованості параметри було розділено на кілька категорій (контроль доступу, шифрування, логування, управління ключами, керування ризиками й управління інцидентами), що охоплюють основні аспекти інформаційної безпеки. Загальний показник конфіденційності системи обчислювався за допомогою зваженої суми, де вагові коефіцієнти уточнювалися залежно від критичності кожного параметра. Для уніфікації шкал і забезпечення коректного подальшого аналізу застосовано методи нормалізації (мінімаксна та Z-нормалізація), завдяки чому отримані значення параметрів можна порівнювати й ефективно інтегрувати в загальну модель. У пропонованому методі для захисту вихідних даних і підвищення приватності використовується диференційна приватність, що забезпечується додаванням випадкового шуму з нормальним розподілом. Такий крок ускладнює процес відновлення початкових показників та мінімізує ризик ідентифікації конкретних записів, зберігаючи при цьому точність сукупних статистичних оцінок. Розроблений підхід містить кілька послідовних етапів: від первинної категоризації й нормалізації даних до реалізації диференційної приватності до аналізу даних у нейронній мережі. Його важливою перевагою є можливість інтегрувати різні аспекти захисту даних у єдину узгоджену систему. Така багатовимірна концепція сприяє гнучкості рішення та дозволяє швидко адаптувати його до оновлених вимог або появи нових загроз. Представлена модель особливо актуальна в галузях, де обробляються чутливі дані: охороні здоров'я, банківському та фінансовому секторах, а також у сфері державного управління й інформаційної безпеки. Запропонований підхід закладає основу для розробки й масштабування безпечних та прозорих систем, які відповідають сучасним стандартам збереження конфіденційності

Ключові слова: захист інформації; безпекові нейронні мережі; нормалізація векторних даних; доступ до інформації; параметри оцінки безпеки; статистичний шум; адаптивне машинне навчання

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 22.08.2025 Revised: 17.11.2025 Accepted: 23.12.2025

UDC 004.415.5 DOI: 10.31649/vitce/3.2025.136

Fuzzy-algorithmic analysis of software reliability

Hanna Rakytyanska

PhD in Technical Sciences, Associate Professor
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0000-0001-5863-3730

Bohdan Prus*

Postgraduate Student
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
https://orcid.org/0009-0008-7214-0949

Abstract. The relevance of the study was due to the need to develop interpretable process-oriented models that allow assessing the growth of the reliability function depending on the distribution of efforts. The aim of the work was to model the processes associated with introducing, detecting, and correcting errors using algorithmic algebra and fuzzy logic. The proposed methodology for software reliability analysis was based on the theory of reliability of algorithmic processes. The logical-algorithmic model of the development process was built on the basis of linear, alternative, and iterative operator structures. The sequence of works without feedback is described by the linear structure. The verification and validation stages were described using alternative and iterative algorithmic structures. The process of checking and correction, when detected errors were immediately removed, and new errors were not introduced, was described by the alternative structure. The debugging process, during which new errors might be introduced, was described by the iterative structure. The logical-algorithmic model in the form of the fuzzy knowledge base made it possible to design software with the required levels of reliability and cost using improving transformations. The system of fuzzy logical equations connected the correctness levels of the working, checking, and correction operations with the possibility of correct execution of the development process. The allocation of efforts was formalised by the improving substitutions introduced in the logical-algorithmic model. The controllable variables associated with improving substitutions were interpreted as the quality of execution of the working, checking, and correction operations. The proposed fuzzy model of software reliability allows to assess the risks of the development process based on expert and experimental information about the reliability and time characteristics of the life cycle stages. The model was constructed by transferring reliable parts of the development process obtained from the histories of errors and defects of previous projects into a process-oriented reliability model of the current project. The example of reliability analysis of the process of developing a mobile application for image aggregation was considered, where the influencing factors are the error-free execution of working, checking, and correction operations. The practical significance of the study lies in the development of a toolset that makes it possible to predict software reliability at different stages of the life cycle, to optimise the allocation of resources between error detection and correction, and to reduce the risks of unsuccessful decisions in design and debugging

Keywords: software reliability; risk assessment of the development process; effort allocation; logical-algorithmic model; fuzzy reliability model

Introduction

The main goal of software development is to provide the best quality product within a limited time (cost). Functional reliability is the main criterion that requires modelling and analysis at various stages of development. As shown in the study D. Hanagal & N. Bhalerao (2021), traditional reliability growth models predict software failures, but

Suggested Citation:
Rakytyanska, H., & Prus, B. (2025). Fuzzy-algorithmic analysis of software reliability. *Information Technologies and Computer Engineering*, 22(3), 136-147. doi: 10.31649/vitce/3.2025.136

*Corresponding author

do not allocate efforts to prevent them. Modelling events that lead to software failures requires the description of the software operation process and the definition of clusters of influential factors (Sokol, 2025). Software reliability analysis requires identifying the structure and parameters of a reliability model obtained through process validation to ensure compliance with requirements. At the same time, as V. Pradhan *et al*. (2023) note, developers need interpretable reliability models that give them insight into where to focus their efforts to reduce the risk of defects. Understanding the interdependence of events and processes is critical to software reliability. Real-world scenarios involve the risk of missing errors in the system because the processes of error detection and removal are imperfect. In the article M. Macak *et al*. (2022), process mining is defined as a set of methods aimed at extracting basic knowledge from processes to interpret data and understand the process in dynamics. The authors of research U. Samal *et al*. (2025) argued that the description of risks and uncertainties inherent in development processes determines the use of fuzzy models of software reliability.

As noted in the study C. Thieme *et al*. (2020), risk assessment is performed by embedding software functional failures and their propagation effects into traditional risk analysis models such as fault trees and event trees. Risk prediction is performed using defect density analysis based on the coordinate model combined with fault information. In practice, the use of combinatorial models such as fault trees or event trees is limited by the complexity of system behaviour. To manage the number of states for software components, Markov chains that model software behaviour using object-oriented modelling templates are proposed by R. Calinescu *et al*. (2021). The transition probabilities between different system states can be obtained from the operating environment. In the study V. Yakovyna & I. Symets (2021), a Markov model of software reliability is proposed that allows predicting the maximum number of operational states using the representation of a higher-order Markov process by an equivalent first-order process with additional virtual states. In a dynamic and uncertain environment, the entropy characterises the probability of transition to the next state of the system. U. Samal & A. Kumar (2024) considered software reliability models with detection and elimination of multiple failures that occur stochastically during testing stages. Reliability models of debugging stages are focused on evaluating the intensity of failures depending on the frequency of errors. In the study X. Chen & Y. Deng (2024), a software risk assessment model is proposed that can measure the uncertainty associated with stochastic testing and debugging processes by introducing belief entropy. Statistical modelling is aimed at identifying and ranking factors associated with the occurrence, detection, and elimination of errors. To consider risk factors, a conceptual model based on expert recommendations on project management is developed by B. Duarte *et al*. (2021). The conceptual model connects influencing factors and their interaction with on-time delivery, where deviations from the schedule are due to the errors in resource allocation. The work S. Butt *et al*. (2023) proposed an ontological model of the software testing process in the form of a knowledge graph, which establishes internal relationships between test cases and defects. Ontological analysis of software systems includes the construction of an ecosystem of software artifacts at different levels of abstraction. In the study H. Ferreira *et al*. (2023), a conceptual reliability models of software-intensive systems using machine learning and cloud technologies is proposed. The model consisting of dozens of concepts and connections between them is aimed at understanding processes and events at the stages of software verification and validation.

Constructing process-based software reliability models is computationally complex. The use of state-based models such as Markov chains is limited by the growth of the state space. The complexity of system behaviour limits the use of conceptual models based on the knowledge graph. At the same time, an approach aimed at transferring reliable components of the previous system to the new one from the failure history of existing systems can be a solution when constructing process-oriented reliability models. Reliable components are formalised as development stages related to error detection and correction. In this case, the model parameters represent a distribution of efforts that is easily interpreted by developers and can be changed according to requirements. The aim of the research was to develop the interpretable process-oriented models of software reliability based on the algebra of algorithms and fuzzy logic.

Materials and Methods

The proposed methodology for software reliability analysis is based on the theory of reliability of algorithmic processes (Rotshtein *et al*., 2007). The uncertainty associated with development stages is described using fuzzy logic (Rotshtein & Rakytyanska, 2012). The principles formulated in the aforementioned studies were used to model the reliability of the software development process. The description of events related to the introduction, detection, and removal of errors is carried out using V.M. Hlushkov's algebra of algorithms (Doroshenko *et al*., 2004). The logical-algorithmic model consists of operators and logical conditions as well as operator and logical structures that describe discrete software development processes. In the theory of reliability of algorithmic processes, a distinction between working operators and correction operators is made. Logical conditions formalise the operations of checking the correctness of the execution of working and modifying operators. Workflows are described using the linear algorithmic structures associated with the introduction of errors. The processes of error detection and elimination are described by the alternative and iterative algorithmic structures. The logical-algorithmic description of the development process is considered as an analogue of the fuzzy knowledge base. Then, the reliability model in the form of the system of fuzzy logical equations connects the possibilities of correct (incorrect) execution of the process and its elements

included in the logical-algorithmic description. Improving transformations embedded in the logical-algorithmic model ensured the increase in the reliability and cost of the development process. Controllable variables associated with improving transformations allowed generating variants of the development process with the required level of quality under time constraints. Software for software reliability analysis was developed in Python.

Logical-algorithmic models of the development process

The software development process was described in the system of V.M. Hlushkov's algorithmic algebras using the typical operator and logical structures (Doroshenko *et al.*, 2004). The linear structure described a sequence of works without checking operations and feedback. Checking operations corresponded to the stages of verification and validation aimed at detecting errors. Correction operations corresponded to the debugging stages aimed at removing errors. The alternative structure described the branched process of "work – checking (testing) – correction without feedback", when errors were detected and immediately removed from the system; the iterative structure described the cyclical process of "work – checking (testing) – correction with feedback", when new errors can be introduced during debugging. The process of developing a null version of software with sequential implementation of stages was described by a linear algorithmic structure of the form:

$$L_0 = A_1 A_2 \ldots A_n, \tag{1}$$

where A_i is the working operator corresponding to the i-th development stage; $i = 1,\ldots, n$; L_0 is the equivalent operator corresponding to the linear structure of the null version.

The software development process is associated with the introduction of errors of the j-type, $j = 1,\ldots, m$, which are subject to detection and removal at the stages of verification and validation. It was assumed that $e_i = \{e_{i1},\ldots, e_{ik_i}\}$ be the set of errors introduced during the execution of the working operator A_i, $k_1 + \cdots + k_n = m$. The processes of verification and validation of the working operator A_i with the detection and removal of errors e_i were described by the following algorithmic structures (Rotshtein *et al.*, 2007; Rotshtein & Rakytyanska, 2012):

- alternative algorithmic structure "work – checking – correction without feedback"

$$B_i = A_i[e_i] \underset{\omega_i}{(} D \vee U_i); \tag{2}$$

- iterative algorithmic structure "work – checking – correction with feedback"

$$C_i = A_i[e_i] \underset{\omega_i}{\{} R_i\}, \tag{3}$$

where ω_i is the logical condition for checking the correctness of the execution of the working operator A_i, $\omega_i = 1(0)$ if the operator A_i is performed correctly (incorrectly); D is the identical operator which is interpreted as fixing the results of checking; U_i is the correction operator during the execution of which detected errors are immediately removed and new errors are not introduced; R_i is the correction operator during the execution of which new errors may be introduced; B_i C_i are the equivalent operators corresponding to the alternative and iterative algorithmic structures.

Software quality management was carried out with the help of improving substitutions introduced into the logical-algorithmic model. Improving substitutions aimed at increasing software reliability are interpreted as the quality of performing working, checking, and correction operations (Rotshtein & Rakytyanska, 2012). Improving substitutions model the allocation of efforts to reduce the risk of introducing errors during the development stage; reducing the risk of missing errors at the testing stage; reducing the risk of leaving errors in the system or introducing new errors at the debugging stage. The distribution of efforts was formalised by adding the following improving substitutions into the logical-algorithmic models (1)-(3):

$$L = (A_1)^{\boldsymbol{X}_1}(A_2)^{\boldsymbol{X}_2}\ldots(A_n)^{\boldsymbol{X}_n}; \tag{4}$$

$$P_i = (A_i)^{\boldsymbol{X}_i} \underset{(\omega_i)^{\boldsymbol{Y}_i}}{(} D \vee (U_i)^{\boldsymbol{Z}_i}); \tag{5}$$

$$Q_i = (A_i)^{\boldsymbol{X}_i} \underset{(\omega_i)^{\boldsymbol{Y}_i}}{\{} (R_i)^{\boldsymbol{Z}_i^*}\}, \tag{6}$$

where $X_i = (x_{i1}, x_{i2},\ldots, x_{ik_i})$ is the vector of parameters that determine the quality of execution of the working operator (A_i) to reduce the risk of introducing the error e_{ij}; $Y_i = (y_{i1}, y_{i2},\ldots, y_{ik_i})$ is the vector of parameters that determine the quality of execution of the checking operator (ω_i) to reduce the risk of missing the error e_{ij} at the stages of verification and validation; $Z_i = (z_{i1}, z_{i2},\ldots, z_{ik_i})$ and $Z_i^* = (z_{i1}^*, z_{i2}^*,\ldots, z_{ik_i}^*)$ are the vectors of parameters that determine the quality of execution of the correction operators (U_i and R_i) to reduce the risk that the error e_{ij} will not be corrected and will remain in the system and new errors will be introduced. The parameters x_{ij}, y_{ij}, z_{ij}, $z_{ij}^* \in \{1, 3, 5, 7, 9\}$, which are chosen in accordance with Saaty's scale (Saaty, 1994; Rotshtein & Rakytyanska, 2012), are interpreted as severity ranks of the errors e_{ij} at the working stages and priority ranks at the stages of checking and correction.

The process of software development, testing, and debugging was described by an equivalent algorithmic structure obtained by replacing sections of the initial algorithm (1) with the improving substitutions (4)-(6). The rules of improving transformations (4)-(6) made it possible to represent the generation of variants of the logical-algorithmic model as inference in a formal grammar (Rotshtein *et al.*, 2007):

$$G = \langle V_t, V_n, L_0, I\rangle, \tag{7}$$

where V_t is the set of operator and logical functional units (terminals), $\{A_i, U_i, R_i\}$ and $\{\omega_i\}$, $i = 1,\ldots, n$; V_n is the set of operator functional structures (nonterminals), $\{B_i, C_i\}$; L_0 is the null version of the logical-algorithmic model; I is the

set of improving substitutions: $\{(A_i)^{X_i}, (U_i)^{Z_i}, (R_i)^{Z_i}\}$ for the working and correction operators; $\{(\omega_i)^{Y_i}\}$ for the logical conditions; $\{L, P_i, Q_i\}$ for the operator structures.

Thus, by selecting improving substitutions, the formal description of the development process was synthesised that ensures acceptable levels of software faultlessness and development time. The functional network corresponding to the null version of the development process is transformed until a logical-algorithmic description is found that satisfies the reliability and cost requirements. The sequential algorithmic structure was chosen as a null option. Each improving substitution meant replacing some operator or logical fragment of the functional network by another fragment with an increased level of reliability at the expense of additional costs.

Fuzzy model of software reliability

The level of correctness of the software development process was assessed using the system of fuzzy rules. An analogue of the fuzzy knowledge base and a carrier of the reliability model is the logical-algorithmic description of events associated with the occurrence, detection, and elimination of the causes of incorrect operation of the software system. The inputs of the process-oriented model are the reliability-time estimates of working, checking, and correction operators. At the output of the process, two classes of situations corresponding to correct (μ^1) and incorrect ($\mu^0 = 1 - \mu^1$) execution of the task are identified. Then, for the sequential algorithmic structure (4) with improving transformations (5), (6), the correctness of completion of the stage A_i, $i = 1,\ldots, n$, is described by a system of fuzzy rules:

IF the working operator A_i is executed correctly

OR errors are correctly detected and removed during checking ω_i and U_i correction ,

AND no new errors have been made,

OR errors are correctly detected and resolved during correction R_i.

(with the possibility of introducing new errors)

THEN the stage is completed correctly.

The following system of fuzzy logical equations that connects the levels of fuzzy correctness of operators and logical conditions with the classes of decisions corresponding to correct (μ^1) and incorrect ($\mu^0 = 1 - \mu^1$) execution of the task is derived from the fuzzy knowledge base:
for the alternative structure

$$\mu_{P_i}^1 = \mu_{A_i}^1(X_i) \cdot \mu_{\omega_i}^1(Y_i) + \mu_{A_i}^0(X_i) \cdot \mu_{U_i}^1(Z_i); \quad (8)$$

for the iterative structure

$$\begin{aligned}\mu_{Q_i}^1 = {} & \mu_{A_i}^1(X_i) \cdot \mu_{\omega_i}^1(Y_i) + \mu_{A_i}^0(X_i) \cdot \mu_{R_i}^1(Z_i^*) + \\ & + \mu_{A_i}^{00} \cdot [\mu_{R_i}^{00} + \mu_{R_i}^{10}] \cdot \mu_{R_i}^1(Z_i^*) + \cdots \\ & + [\mu_{A_i}^{00}]^q \cdot [\mu_{R_i}^{00} + \mu_{R_i}^{10}]^q \cdot \mu_{R_i}^1(Z_i^*), \end{aligned} \quad (9)$$

where the risk of repeated correction is

$$\mu_{A_i}^{00} = 1 - [\mu_{A_i}^1(X_i) \cdot \mu_{\omega_i}^1(Y_i) + \mu_{A_i}^0(X_i) \cdot \mu_{R_i}^1(Z_i^*)], \quad (10)$$

where $\mu_{A_i}^1$ ($\mu_{A_i}^0$), $\mu_{\omega_i}^1$ ($\mu_{\omega_i}^0$), and $\mu_{U_i}^1$ ($\mu_{U_i}^0$) are the possibilities of correct (incorrect) execution of the working (A_i), checking (ω_i), and correction (U_i) operators; $\mu_{R_i}^1(\mu_{R_i}^{00}, \mu_{R_i}^{10})$ is the distribution of fuzzy correctness of the cyclic correction operator R_i; q is the number of verification and correction cycles for the iterative structure; $\mu_{P_i}^1$ ($\mu_{P_i}^0$) and $\mu_{Q_i}^1$ ($\mu_{Q_i}^0$) is the possibility of correct (incorrect) execution of the equivalent operators P_i and Q_i.

The reliability of working, checking, and correcting operators was modelled on the basis of expert and experimental data on the distribution of errors of various types (interface errors, errors in the logic of the program, errors in the processing of parallel data flows, etc.). Following the m-ary concept of errors (Rotshtein *et al.*, 2007), the possibility of identifying and correcting errors of the j-th type, $j = 1,\ldots, k_i$, was considered at the i-th stage. Managing the risks of introducing and omitting errors when executing working, checking, and correction operators was carried out using controllable variables as follows:

$$\begin{aligned}\mu_{A_i}^1(X_i) &= \prod_{j=1}^{k_i}\left(1 - [\mu_{A_{ij}}^0]^{x_{ij}}\right), \\ \mu_{\omega_i}^1(Y_i) &= \prod_{j=1}^{k_i}\left(1 - [\mu_{\omega_{ij}}^0]^{y_{ij}}\right), \\ \mu_{U_i}^1(Z_i) &= \prod_{j=1}^{k_i}\left(1 - [\mu_{U_{ij}}^0]^{z_{ij}}\right),\end{aligned} \quad (11)$$

$$\mu_{R_i}^1(Z_i^*) = \prod_{j=1}^{k_i}\left(1 - [\mu_{R_{ij}}^{00} + \mu_{R_{ij}}^{10}]^{z_{ij}^*}\right), \quad (12)$$

where $\mu_{A_i}^1$ ($\mu_{A_i}^{0L}$) is the possibility of avoiding (introducing) the error e_{ij} when executing the working operator A_i; $\mu_{\omega ij}^1(\mu_{\omega ij}^0)$ is the possibility of detecting (missing) the error e_{ij} when checking the truth of the logical condition ω_i; $\mu_{Uij}^1(\mu_{Uij}^0)$ is the possibility of removing (leaving in the system) the error e_{ij} when executing the correction operator U_i; $\mu_{Rij}^1(\mu_{Rij}^{00}, \mu_{Rij}^{10})$ is the possibility of removing (leaving in the system) the error e_{ij} and introducing new errors when executing the correction operator R_i.

Fuzzy correctness of the multistage development process described by a sequence of the alternative and iterative algorithmic structures was defines as follows:

$$\mu^1 = \prod_{i=1}^{n}(\mu_{S_i}^1(X_i, Y_i, Z_i, Z_i^*)), S_i \in \{P_i, Q_i\}. \quad (13)$$

The execution time of the working (t_{A_i}), checking (t_{ω_i}), and correction (t_{U_i}, t_{R_i}) operations was calculated in proportion to the development time of the null version of the working operator t_{Ai}^0 as follows:

$$\begin{aligned}t_{A_i}(X_i) &= (1 + \frac{\max_j(x_{ij})}{x_{max}})t_{A_i}^0, t_{\omega_i}(Y_i) = \frac{\max_j(y_{ij})}{y_{max}}t_{A_i}^0, \\ t_{U_i(R_i)}(Z_i) &= \frac{\max_j(z_{ij})}{z_{max}}t_{A_i}^0,\end{aligned} \quad (14)$$

where the maximum quality ratings are $x_{max}, y_{max}, z_{max} = 9$.

The execution time of the algorithms (4)-(6) was estimated as follows:
for the alternative structure

$$t_P^* = \sum_{i=1}^{n}(t_{A_i}(X_i) + t_{\omega_i}(Y_i) + t_{U_i}(Z_i)); \quad (15)$$

for the iterative structure

$$t_Q^* = \sum_{i=1}^{n}(t_{A_i}(X_i) + q(t_{\omega_i}(Y_i) + t_{R_i}(Z_i^*))), \quad (16)$$

where t_P^*, t_Q^* is the execution time of the equivalent operators corresponding to the alternative and iterative structures.

Thus, correlations (8)-(13) define the fuzzy model of software reliability growth. The increase in software reliability was achieved by managing risks of incorrect performance of the task. Parameters of working, checking, and correction operators determine the efforts aimed at reducing the risks of errors at all stages of development. The growth of the reliability function is ensured by introducing the quality indicators of working, checking, and correction operations in accordance with the Saaty's scale. Fuzzy logical equations model the increase in the level of fuzzy correctness depending on the distributed efforts. The development time determined by correlations (15), (16) is calculated in accordance with the priority ranks of working, checking, and correction operators.

Results and Discussion

Example: Risk assessment of the mobile application development.

The problem of reliability analysis of a mobile application for image aggregation is considered (Rakytyanska & Prus 2024). It is assumed that a logical-algorithmic model of the development process is given, where risk assessments of the working, checking, and correction operations can be obtained on the basis of already completed projects. The problem of reliability analysis was formulated as follows. For the given logical-algorithmic model, it is necessary to minimise risks of introducing, missing, and leaving errors in the system by distributing efforts to perform working, checking, and correction operations. The development time should not exceed 16 weeks.

The logical-algorithmic model of the development process looks as follows:

$$A^* = A_0\,[e_0]_{\omega_0}(D \vee U_0)\ldots A_2\,[e_2]_{\omega_2}(D \vee U_2)\,A_3\,[e_3]_{\omega_3}\{R_3\}\,A_4\,[e_4]_{\omega_4}(D \vee U_4)$$
$$A_5\,[e_5]_{\omega_5}(D \vee U_5)\,A_6\,[e_6]_{\omega_6}\{R_6\}\ldots A_{15}\,[e_{15}]_{\omega_{15}}\{R_{15}\}$$
$$A_{16}\,[e_{16}]_{\omega_{16}}(D \vee U_{16})\ldots A_{19}\,[e_{19}]_{\omega_{19}}(D \vee U_{19}). \quad (17)$$

Here A_0 – requirements analysis; $e_0 = \{e_{0.1}\}$, where $e_{0.1}$ – requirements incorrectly defined by the customer;

A_1 – development of technical specifications for software design; $e_1 = \{e_{1.1},\ldots, e_{1.5}\}$, where $e_{1.1}$ – incorrect formulation of technical requirements; $e_{1.2}$ – lack of clear software goals; $e_{1.3}$ – improper distribution of roles and responsibilities; $e_{1.4}$ – changing requirements; $e_{1.5}$ – incomplete detailing of the project;

A_2 – development of technical specifications for programming; $e_2 = \{e_{2.1},\ldots, e_{2.4}\}$, where $e_{2.1}$ – misunderstanding of technical aspects; $e_{2.2}$ – inconsistency of resources and deadlines; $e_{2.3}$ – lack of quality control; $e_{2.4}$ – failure to consider risks;

A_3 – development of UI/UX design; $e_3 = \{e_{3.1},\ldots, e_{3.7}\}$, where $e_{3.1}$ – inconsistency with user needs; $e_{3.2}$ – complex and confusing interface; $e_{3.3}$ – poor readability and contrast; $e_{3.4}$ – non-optimised images; $e_{3.5}$ – lack of processing of user paths; $e_{3.6}$ – ignoring updates and trends; $e_{3.7}$ – non-adaptive design;

A_4 – analysis of implementation methods; $e_4 = \{e_{4.1}, e_{4.2}\}$, where $e_{4.1}$ – neglect of resources; $e_{4.2}$ – neglecting expansion needs;

A_5 – analysis of architectural solutions; $e_5 = \{e_{5.1}, e_{5.2}\}$, where $_{5.1}$ – disregarding software functionality; $e_{5.2}$ – neglecting expansion needs;

A_6 – software architecture development; $e_6 = \{e_{6.1},\ldots, e_{6.3}\}$, where $e_{6.1}$ – errors in the analysis of architectural solutions; $e_{6.2}$ – suboptimal performance; $e_{6.3}$ – dissatisfaction with security needs;

A_7 – interface development; $e_7 = \{e_{7.1},\ldots, e_{7.5}\}$, where $e_{7.1}$ – design pattern mismatch; $e_{7.2}$ – using the wrong style elements; $e_{7.3}$ – inappropriate animations; $e_{7.4}$ – non-adaptive interface; $e_{7.5}$ – errors in the text;

A_8 – development of software modules; $e_8 = \{e_{8.1},\ldots, e_{8.4}\}$, where $e_{8.1}$ – errors in the program code; $e_{8.2}$ – third-party library errors; $e_{8.3}$ – errors in the work algorithms of software modules; $e_{8.4}$ – unoptimised code;

A_9 – development of image processing module; $e_9 = \{e_{9.1},\ldots, e_{9.4}\}$, where $e_{9.1}$ – processing algorithm errors; $e_{9.2}$ – internal library error; $e_{9.3}$ – hardware processing algorithms are not supported; $e_{9.4}$ – overloading of the computing resources of the device;

A_{10} – database development; $e_{10} = \{e_{10.1},\ldots, e_{10.3}\}$, where $e_{10.1}$ – database normalisation is broken; $e_{10.2}$ – wrong data types; $e_{10.3}$ – incorrect types of relationships between tables;

A_{11} – connecting logic to the interface; $e_{11} = \{e_{11.1}, e_{11.2}\}$, where $e_{11.1}$ – inconsistency of the interface with the program logic; $e_{11.2}$ – interface design errors;

A_{12} – software testing; $e_{12} = \{e_{12.1},\ldots, e_{12.5}\}$, where $e_{12.1}$ – using the wrong user flow during testing; $e_{12.2}$ – there is no verification of software operation on different devices; $e_{12.3}$ – no smoke testing; $e_{12.4}$ – there is no verification of the problem in several approaches; $e_{12.5}$ – no testing with poor or no internet connection;

A_{13} – error correction; $e_{13} = \{e_{13.1},\ldots, e_{13.4}\}$, where $e_{13.1}$ – incorrectly formulated ways of reproducing the error; $e_{13.2}$ – not following all paths to reproduce the error; $e_{13.3}$ – no playback error with correct playback paths; $e_{13.4}$ – there is no verification of the entire software module when the software code changes in it;

A_{14} – checking the requirements of the technical task; $e_{14} = \{e_{14.1}, e_{14.2}\}$, where $e_{14.1}$ – inconsistency of the current design with the requirements; $e_{14.2}$ – non-compliance of the software functionality with the requirements of the technical task;

A_{15} – verification of the software product by the customer; $e_{15} = \{e_{15.1}, e_{15.2}\}$, where $e_{15.1}$ – non-compliance of the software with the primary requirements; $e_{15.2}$ – non-compliance of the interface with the primary requirements;

A_{16} – preparation for publication; $e_{16} = \{e_{16.1},\ldots, e_{16.4}\}$, where $e_{16.1}$ – there is no full testing of the software before publication; $e_{16.2}$ – uncorrected critical software errors;

$e_{16.3}$ – errors in the description of the user terms of service provision; $e_{16.4}$ – there is no testing of alpha and beta versions of the software;

A_{17} – checking requirements before publishing to AppStore, Google Play; $e_{17} = \{e_{17.1},..., e_{17.4}\}$, where $e_{17.1}$ – non-compliance with the requirements of application stores; $e_{17.2}$ – misrepresentation of user data; $e_{17.3}$ – non-compliance with the requirements to be effective after publication; $e_{17.4}$ – non-compliance with country-specific publication requirements;

A_{18} – publication; $e_{18} = \{e_{18.1}, e_{18.2}\}$, where $e_{18.1}$ – error when uploading the application assembly to the application store; $e_{18.2}$ – publication of the application without taking into account analytical data;

A_{19} – support; $e_{19} = \{e_{19.1},..., e_{19.3}\}$, where $e_{19.1}$ – no resources for support; $e_{19.2}$ – critical bug fixes are missing; $e_{19.3}$ – no support for new platform requirements.

In (16), the stages $A_0 - A_2, A_4, A_5, A_{16} - A_{19}$ are described by the alternative algorithmic structures without feedback, and the stages A_3, $A_{16} - A_{19}$ are described by the iterative algorithmic structures with feedback. Risk assessments for the working, checking, and correction operators are given in Table 1. Error distribution ranges were obtained on the basis of expert and experimental data. To estimate the fuzzy correctness of the operators and operator structures in (16), the histories of errors and defects of 56 already completed projects in the field of mobile application development were considered. The reliability of working operators ($\mu_{A_i}^1$) is determined based on fuzzy estimates of the frequency of introducing errors of various types. The reliability of checking ($\mu_{\omega i}^1$) and correction (μ_{Ui}^1, μ_{Ri}^1) operators is determined based on fuzzy estimates of the frequency of detection and correction of errors of various types. Given the fuzzy correctness (μ^1), the risk ($\mu^0 = 1 - \mu^1$) is calculated.

Table 1. Risk assessments for the working, checking, and correction operators

Stage	Working operator	Checking operator	Correction without feedback	Correction with feedback	
A_i	μ_{Aij}^0	$\mu_{\omega ij}^0$	μ_{Uij}^0	μ_{Rij}^{00}	μ_{Rij}^{10}
A_0	0.392	0.464	0.214	-	-
A_1	0.115-0.428	0.185-0.396	0.120-0.275	-	-
A_2	0.196-0.410	0.142-0.339	0.107-0.250	-	-
A_3	0.250-0.375	0.160-0.285	-	0.071-0.160	0.053-0.107
A_4	0.160-0.267	0.178-0.250	0.089-0.178	-	-
A_5	0.214-0.357	0.196-0.304	0.185-0.214	-	-
A_6	0.304-0.410	0.214-0.392	-	0.089-0.178	0.071-0.142
A_7	0.232-0.446	0.285-0.464	-	0.107-0.267	0.053-0.125
A_8	0.125-0.392	0.196-0.428	-	0.160-0.250	0.125-0.178
A_9	0.214-0.464	0.125-0.375	-	0.142-0.232	0.107-0.160
A_{10}	0.107-0.250	0.196-0.267	-	0.089-0.214	0.071-0.125
A_{11}	0.285-0.392	0.178-0.304	-	0.178-0.232	0.089-0.160
A_{12}	0.196-0.428	0.214-0.339	-	0.160-0.304	0.107-0.142
A_{13}	0.304-0.446	0.267-0.410	-	0.125-0.160	0.071-0.125
A_{14}	0.107-0.160	0.089-0.125	-	0.071-0.089	0.053-0.071
A_{15}	0.125-0.178	0.107-0.160	-	0.107-0.214	0.053-0.089
A_{16}	0.089-0.232	0.125-0.178	0.071-0.160	-	-
A_{17}	0.196-0.267	0.160-0.214	0.107-0.178	-	-
A_{18}	0.160-0.214	0.142-0.196	0.053-0.071	-	-
A_{19}	0.107-0.178	0.125-0.214	0.089-0.142	-	-

Source: created by the authors

When calculating the reliability of algorithmic structures using formulas (8), (9), the risk of incorrect completion of development stages for preliminary risk assessments is 10-25%. This means that without proper resource allocation, the possibility of cascading and interactive effects as the consequences of errors propagation steadily increases. As a result, the goals of the project cannot be achieved within the established time frame. Assessments of the correctness of software development stages after efforts distribution are given in Table 2. Risk management is carried out based on formulas (11), (12) by dividing efforts (x_{ij}, y_{ij}, z_{ij}) to reduce the risks of introduction (μ_{Aij}^0), imperfect detection ($\mu_{\omega ij}^0$), and incomplete correction (μ_{Uij}^0, μ_{Rij}^0) of the set of errors $e_i = \{e_{ij}\}$, $j = 1,..., k_i$. The result of the effort distribution is a stable growth of the reliability function due to the correct execution of the working (μ_{Ai}^1), checking ($\mu_{\omega i}^1$), and correction (μ_{Ui}^1, μ_{Ri}^1) operators. Estimates of the correct completion of each stage (μ_{Pi}^1, μ_{Qi}^1) associated with correctness of the execution of the alternative or iterative structure, are obtained using formulas (8), (9).

Table 2. Fuzzy correctness of development stages after distribution of efforts among working, checking, and correction operators

Stage	Working operator		Checking operator		Correction without feedback			Correction with feedback		
A_i	x_{ij}	$\mu^1_{A_i}$	y_{ij}	$\mu^1_{\omega_i}$	z_{ij}	$\mu^1_{U_i}$	$\mu^1_{P_i}$	z^*_{ij}	$\mu^1_{R_i}$	$\mu^1_{Q_i}$
A_0	5	0.940	7	0.995	7	0.998	0.996	-	-	-
A_1	5-7	0.986	7-8	0.998	6-8	0.999	0.998	-	-	-
A_2	4-5	0.972	6-7	0.998	6-7	0.997	0.996	-	-	-
A_3	4-5	0.980	5-6	0.998	-	-	-	5-6	0.997	0.995
A_4	3-4	0.981	4-5	0.997	3-4	0.998	0.996	-	-	-
A_5	4-5	0.984	5-6	0.997	5-6	0.995	0.997	-	-	-
A_6	5-6	0.978	6-7	0.996	-	-	-	6-7	0.997	0.996
A_7	5-6	0.963	5-7	0.996	-	-	-	7-8	0.999	0.996
A_8	4-6	0.974	6-7	0.997	-	-	-	7-8	0.996	0.995
A_9	4-5	0.950	5-6	0.996	-	-	-	6-7	0.995	0.992
A_{10}	3-4	0.984	4-5	0.998	-	-	-	5-6	0.999	0.997
A_{11}	3-4	0.973	4-5	0.995	-	-	-	4-5	0.996	0.995
A_{12}	5-6	0.936	6-8	0.998	-	-	-	7-8	0.996	0.994
A_{13}	5-6	0.990	6-7	0.996	-	-	-	7-8	0.998	0.996
A_{14}	3-4	0.982	4-5	0.998	-	-	-	4-5	0.997	0.996
A_{15}	3-4	0.990	4-5	0.997	-	-	-	4-5	0.998	0.995
A_{16}	2-3	0.947	2-3	0.995	2-3	0.996	0.995	-	-	-
A_{17}	2-3	0.981	3-4	0.998	2-3	0.994	0.997	-	-	-
A_{18}	2-3	0.954	3-4	0.997	3-4	0.998	0.996	-	-	-
A_{19}	2-3	0.968	4-3	0.998	2-3	0.997	0.995	-	-	-

Source: created by the authors

The obtained results demonstrate the correct ranking of risks during the performance of working, verification, and correction operations. The possibility of correct completion of the project is estimated using formula (13). Risk assessments during the analysis of requirements and development of technical specifications at the stages A_0-A_2 amount to approximately 1%. Risk assessments during the analysis of user's design, implementation methods, and architectural solutions at the stages A_3-A_5 are less than 1.5%. Risk assessments during the development and testing at the stages A_6-A_{15} are around 5%. Risk assessments in preparation for publication and support at the stages A_{16}-A_{19} are less than 2%. Reducing the risk of incorrect completion of the previous stage A_{i-1} allows avoiding the effects of cascading and interaction when spreading the consequences of errors $e_i = \{e_{ij}\}, j = 1,\ldots, k_i$, at the stages $A_i, A_{i+1},\ldots, A_n$.

The schedule for the execution of the sequential discrete process (16) is shown in Figure 1. The calendar plan was calculated using formulas (14), (15) taking into account the costs of checking and correction in accordance with the distribution of efforts. To comply with time constraints, the start of each stage was chosen as early as possible. The development of technical specifications at the stages A_0-A_2 takes 1-2 weeks. The development of user's design and architectural solutions at the stages A_3-A_5 takes approximately 4 weeks. The duration of the development and testing stages A_6-A_{15} is around 12 weeks. The final stages of preparation for publication A_{16}-A_{19} take no more than 2 weeks. When calculating the calendar plan, contextual histories of already completed projects were used, where the duration of the development and testing stages makes up 80% of the project time. The time spent on detecting and correcting errors was estimated in proportion to the development time of the null version of the stage, depending on the priority of the errors. As a result of proper distribution of efforts, the software development period does not exceed 16 weeks, which guarantees timely delivery.

Within the given schedule, the progress of the project in the form of the cumulative flow diagram can be predicted using the system of fuzzy rules and assessments of correct execution of the operators and logical conditions (Tables 1, 2). The number of tasks in development and testing is known for each week of the calendar plan. The decision to accept the task (or return it for revision) is made based on the possibility of correct execution of the working and checking operators $\mu^1_{A_i}$, $\mu^1_{\omega_i}$, The decision about the successful revision is made based on the possibility of correct execution of the debugging operators $\mu^1_{U_i}$, $\mu^1_{R_i}$. The decision about the successful completion of the stage is made based on the possibility of correct execution of the alternative or iterative structure $\mu^1_{P_i}$, $\mu^1_{Q_i}$ that allows to estimate the number of completed tasks and the number of tasks in the queue. The experimental and model cumulative flow diagrams are shown in Figure 2.

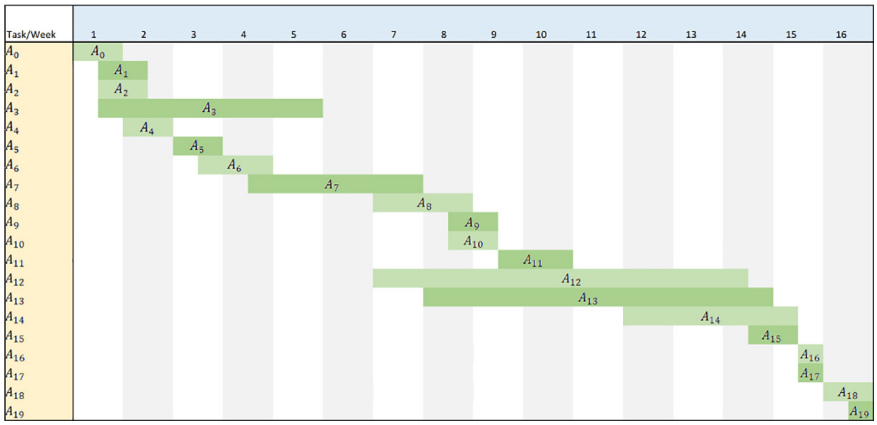


Figure 1. Development process schedule

Source: created by the authors

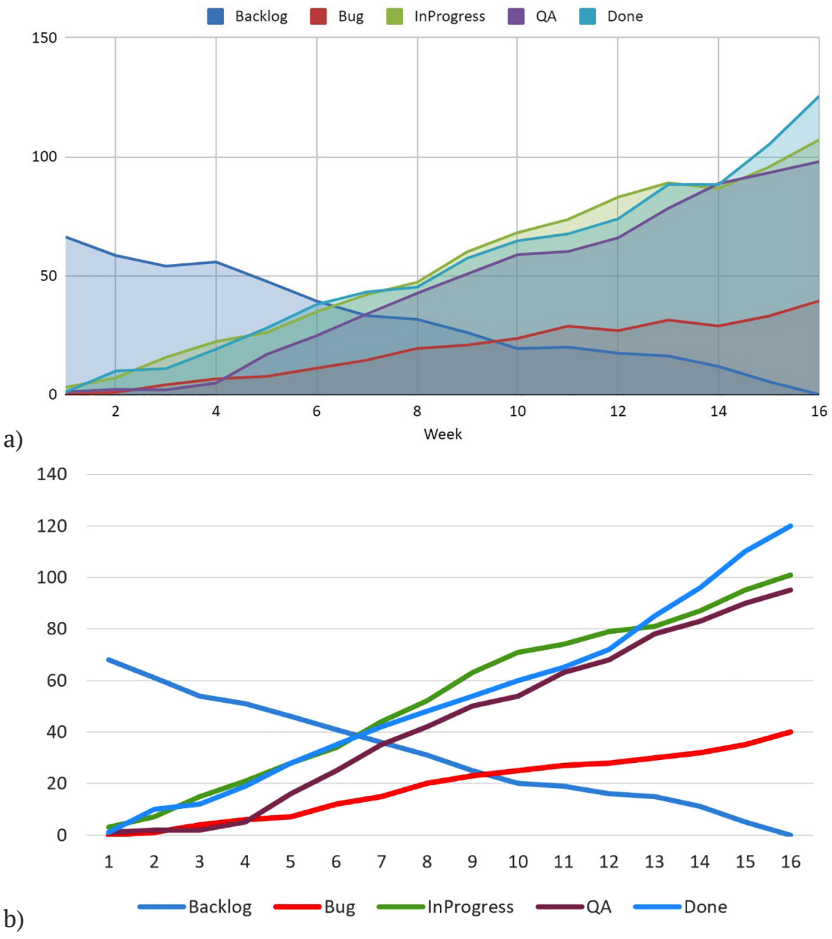


Figure 2. Experimental (a) and model (b) cumulative flow diagram

Source: created by the authors

Comparison of the experimental and model dynamics of the project indicates stable software quality management. As a result of the correct distribution of efforts, the number of completed tasks is steadily increasing; the number of tasks in development and testing remains stable; the number of tasks in the queue is steadily decreasing. In practice, dynamics of a real project allows for short-term horizontal sections indicating the resolution of problems. In case of periods of decline, the model can predict the recovery of the chart. Thus, the proposed model demonstrates the ability to correctly approximate the progress of real projects under conditions of risks and uncertainty.

Software developed in Python for software reliability analysis is an intelligent system that implements a

logical-algorithmic model of the development process. The system is integrated with the analytical module by forming fuzzy rules that take into account the possibility of occurring, detecting, and correcting errors at different stages of the software life cycle. The system allows modelling the processes "work – checking – correction" using improving substitutions and assessing the level of correctness based on a system of fuzzy logical equations. The controllable variables embedded in the logical-algorithmic model formalise the distribution of efforts according to task priority. Visualisation tools make it possible to analyse the dynamics of the project depending on the distributed efforts. For practical application, the model is integrated with repositories of error and defect histories, automatically adapting the parameters of fuzzy rules based on empirical data from previous projects.

Discussion of the results of evaluating the effectiveness of the software reliability model

This paper proposes a software reliability model that allows generating improving transformations of the development process in the form of linguistic rules to prevent the risks of software defects. Such rules are the carrier of the software quality management model, as they give practitioners the opportunity to distribute efforts under limited time conditions. The logical-algorithmic model of software reliability is obtained on the basis of an intelligent analysis of software development processes. The process-oriented model predicts the risks of software defects based on assessments of the correct execution of the process elements, such as the stages of development, testing, and debugging. The principal difference is the integration of the controllable variables into the logical-algorithmic description of the software life circle that allows managing the development risks. Improving substitutions allow to simulate events that ensure software reliability growth at the development stages. As a result, the proposed approach, which is similar to knowledge distillation, allows transferring reliable parts of the previous projects into a process-oriented reliability model of the current project.

Distillation-based models of software reliability aim to define action plans based on software analytics and practitioner findings. In the study B. Littlewood *et al*. (2020), failure rate data of existing software is used as a prior distribution when assessing the reliability of a new system. This approach ensures transferring reliable components of the previous system to the new one from the history of failures of existing systems. The article A. Filippetto *et al*. (2021) consider a database of failures and defects of already completed projects as a context history. The failure history of existing software is used to assess the reliability of a new system by modelling scenarios using distance-based similarity measures. In the work M. Asif & J. Ahmed (2020), a decision support system is developed that automatically generates rules to reduce software risks based on frequent failure patterns. A rule-based machine learning approach is used to establish relationships between risk factors and software reliability, where the previous failure cases and the corresponding action plan are associated with rules. N. Alnahdi & R. Alnanih (2024) consider an information model that integrates usability testing and reliability analysis at the stage of interface design to enhance the system performance based on user experience. The conceptual model is adjusted to identify risks using expert recommendations for minimising risks observed during the implementation of similar projects. The work D. Rajapaksha *et al*. (2022) proposes a defect prediction model for the automatic development of planning strategies using a qualitative study and empirical assessment of the impact factors determined by the software development methodology. Current planning actions are generated in the form of rules-based explanations and associated risk thresholds. The rule-based model proposed by T. Hovorushchenko (2021) predicts the risks of software defects based on many factors obtained from the analysis of experimental data. The risk management model generates planning strategies directly in the form of expert recommendations. R. Ouriques *et al*. (2023) use the approach based on the grounded theory to generate explanations regarding the compliance of the current state of the project with the requirements. The intelligent process analysis allows reconstructing the sequence of events that can cause software failure. As noted in the study J. Díaz *et al*. (2023), modelling condition-event relationships using the grounded theory requires transparency and replicability, which improve the trustworthiness of the generated recommendations. Qualitative data analysis in empirical software research takes into account the perspectives of a group of experts to structure codified knowledge based on a consistent interpretation of events.

Unlike process-based models that examine clusters of factors influencing condition-event relationships, the proposed model uses a time series approach. Correction of the execution of the multidimensional discrete process with the m-ary concept of errors is carried out by selecting improving substitutions. The number of model parameters is reduced to the number of error types k_i at each stage A_i, $i = 1,\ldots, n$. Then, the following groups of risks are considered: risks of introducing errors during development; risks of missing errors during testing; risks of leaving errors in the system or introducing new errors due to imperfect debugging. In particular, to assess the risks of the mobile application development process, from 2 to 7 risk factors or types of errors were considered at each stage of the life cycle $A_0, \ldots, A_{19}$. Therefore, in order to manage the quality of development, it is sufficient to distribute efforts for the correct completion of the stages, where the processes of introducing, detecting, and removing of errors are described by the algorithmic structures "work – checking – correction". Improving transformations formalised by the controllable variables are associated with the thoroughness of development due to the increase of the working time or the skills of developers. The application of the proposed model is limited to discrete algorithmic processes, where risk management is carried out within the time frame of each stage.

Conclusions

This article proposes an approach to modelling the reliability of the software development process based on the algebra of algorithms and fuzzy logic. The multidimensional discrete process of the software development is described using the modified system of V.M. Hlushkov's algorithmic algebras. The algorithmic description of events related to the introduction, detection, and removal of errors is considered as the fuzzy knowledge base "work – checking – correction". The linear structure describes a sequence of works without feedback. The alternative structure describes the process of testing and correction where errors are detected and immediately removed from the system. The iterative structure describes the debugging process with the possibility of introducing new errors. The logical-algorithmic model makes it possible to develop algorithms with the required levels of correctness and cost based on expert and experimental reliability assessments obtained at the stages of the software life cycle.

A fuzzy model of software reliability growth is proposed. The reliability model in the form of the system of fuzzy logical equations connects the possibility of correct (incorrect) execution of the process and the assessments of correctness of the working, checking, and correction operators. Software reliability analysis is associated with assessing the risks that arise during development, verification, and validation due to non-compliance with design requirements. Risk management is carried out with the help of improving substitutions embedded into the logical-algorithmic model. Improving substitutions allow modelling the distribution of efforts to reduce the risk of introducing errors during the development stage; the risk of missing errors at the testing stage; the risk of leaving errors in the system or introducing new errors at the debugging stage. The growth of the reliability function and the progress of the project are ensured by the introduction of indicators of the quality of execution of operators and logical conditions in accordance with Saaty's scale. The synthesis of the fuzzy knowledge base that ensures acceptable levels of software risks and development time is carried out by selecting controllable variables associated with improving substitutions.

Further research involves training the reliability model on experimental data in the form of histories of errors and defects. This approach consists in building and training membership functions of fuzzy correctness for the operators and logical conditions as well as rule weights for the algorithmic structures. A model trained by transferring reliable elements of the development process from the training dataset to the logic-algorithmic model will make it possible to predict the project dynamics depending on the distributed efforts.

Acknowledgements

None.

Funding

The study was not funded.

Conflict of Interest

None.

References

[1] Alnahdi, N., & Alnanih, R. (2024). A novel information model for software interface reliability in the software development life cycle. *Procedia Computer Science*, 251, 116-123. doi: 10.1016/j.procs.2024.11.091.

[2] Asif, M., & Ahmed, J. (2020). A novel case base reasoning and frequent pattern based decision support system for mitigating software risk factors. *IEEE Access*, 8, 102278-102291. doi: 10.1109/ACCESS.2020.2999036.

[3] Butt, S., Ur Rehman Khan, S., Hussain, S., & Wang, W.-L. (2023). A conceptual model supporting decision-making for test automation in Agile-based Software Development. *Data & Knowledge Engineering*, 144, article number 102111. doi: 10.1016/j.datak.2022.102111.

[4] Calinescu, R., Paterson, C., & Johnson, K. (2021). Efficient parametric model checking using domain knowledge. *IEEE Transactions on Software Engineering*, 47(6), 1114-1133. doi: 10.1109/TSE.2019.2912958.

[5] Chen, X., & Deng, Y. (2024). Evidential software risk assessment model on ordered frame of discernment. *Expert Systems with Applications*, 250, article number 123786. doi: 10.1016/j.eswa.2024.123786.

[6] Díaz, J., Pérez, J., Gallardo, C., & González-Prieto, A. (2023). Applying inter-rater reliability and agreement in collaborative Grounded Theory studies in software engineering. *Journal of Systems and Software,* 195, article number 111520. doi: 10.1016/j.jss.2022.111520.

[7] Doroshenko, A., Finin, G., & Tceitlin, G. (2004). *Algebra-algorithmic basics of programming*. Kyiv: Naukova Dumka.

[8] Duarte, B., de Almeida Falbo, R., Guizzardi, G., Guizzardi, R., & Silva Souza, V.E. (2021). An ontological analysis of software system anomalies and their associated risks. *Data & Knowledge Engineering*, 134, article number 101892. doi: 10.1016/j.datak.2021.101892.

[9] Ferreira, H., Nakagawa, Y., & Santos, P. (2023). Towards an understanding of reliability of software-intensive systems-of-systems. *Information and Software Technology,* 158, article number 107186. doi: 10.1016/j.infsof.2023.107186.

[10] Filippetto, A., Lima, R., Luis, J., & Barbosa, V. (2021). A risk prediction model for software project management based on similarity analysis of context histories. *Information and Software Technology,* 131, article number 106497. doi: 10.1016/j.infsof.2020.106497.

[11] Hanagal, D., & Bhalerao, N. (2021). *Software reliability growth models*. Singapore: Springer.

[12] Hovorushchenko, T. (2021). Method of the software risks management. In *Proceedings of the 2nd international workshop on computational & information technologies for risk-informed systems* (pp. 26-38). Kherson: CITR.
[13] Littlewood, B., Salako, K., Strigini, L., & Zhao, X. (2020). On reliability assessment when a software-based system is replaced by a thought-to-be-better one. *Reliability Engineering & System Safety*, 197, article number 106752. doi: 10.1016/j.ress.2019.106752.
[14] Macak, M., Daubner, L., Sani, F., & Buhnova, B. (2022). Process mining usage in cybersecurity and software reliability analysis: A systematic literature review. *Array*, 13, article number 100120. doi: 10.1016/j.array.2021.100120.
[15] Ouriques, R., Wnuk, K., Gorschek, T., & Svensson, B. (2023). The role of knowledge-based resources in Agile Software Development contexts. *Journal of Systems and Software*, 197, article number 111572. doi: 10.1016/j.jss.2022.111572.
[16] Pradhan, V., Kumar, A., & Dhar, J. (2023). Emerging trends and future directions in software reliability growth modeling. In H. Garg & M. Ram (Eds.), *Advances in reliability science, engineering reliability and risk assessment* (pp. 131-144). Amsterdam: Elsevier. doi: 10.1016/B978-0-323-91943-2.00011-3.
[17] Rajapaksha, D., Tantithamthavorn, C., Jiarpakdee, J., Bergmeir, C., Grundy, J., & Buntine, W. (2022). SQAPlanner: Generating data-informed software quality improvement plans. *IEEE Transactions on Software Engineering,* 48(8), 2814-2835. doi: 10.1109/TSE.2021.3070559.
[18] Rakytyanska, H., & Prus, B. (2024). Constructing prototype-based granular fuzzy rules for scene classification on mobile devices. In S. Babichev & V. Lytvynenko (Eds.), *Lecture notes in data engineering, computational intelligence, and decision-making* (pp. 194-218). Cham: Springer. doi: 10.1007/978-3-031-70959-3_10.
[19] Rotshtein, A., & Rakytyanska, H. (2012). *Fuzzy evidence in identification, forecasting and diagnosis.* Heidelberg: Springer.
[20] Rotshtein, A., Shtovba, S., & Kozachko, A. (2007). *Modeling and optimization of multivariable algorithmic processes reliability*. Vinnytsia: UNIVERSUM.
[21] Saaty, T.L. (1994). *Fundamentals of decision making and priority theory with the analytic hierarchy process*. Pittsburgh: RWS Publications.
[22] Samal, U., & Kumar, A. (2024). A software reliability model incorporating fault removal efficiency and it's release policy. *Computational Statistics,* 39, 3137-3155. doi: 10.1007/s00180-023-01430-9.
[23] Samal, U., Kushwaha, S., Nain, G., Singh, S., Usmani, S., & Kumar, A. (2025). Chapter 14. Necessity of fuzzy logic: Trends in software reliability assessment. In A. Kumar, A.S. Bhandari & M. Ram (Eds.), *Reliability assessment and optimization of complex systems* (pp. 275-285). Amsterdam: Elsevier. doi: 10.1016/B978-0-443-29112-8.00009-8.
[24] Sokol, O. (2025). Automation of error detection in code using machine learning. *Bulletin of Cherkasy State Technological University*, 30(1), 35-47. doi: 10.62660/bcstu/1.2025.35.
[25] Thieme, C., Mosleh, A., Utne, I., & Hegde, J. (2020). Incorporating software failure in risk analysis. Part 1: Software functional failure mode classification. *Reliability Engineering & System Safety*, 197, article number 106803. doi: 10.1016/j.ress.2020.106803.
[26] Yakovyna, V., & Symets, I. (2021). Reliability assessment of CubeSat nanosatellites flight software by high-order Markov chains. *Procedia Computer Science*, 192(C), 447-456. doi: 10.1016/j.procs.2021.08.046.

Нечіткий алгоритмічний аналіз надійності програмного забезпечення

Ганна Ракитянська

Кандидат технічних наук, доцент
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0000-0001-5863-3730

Богдан Прус

Аспірант
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
https://orcid.org/0009-0008-7214-0949

Анотація. Актуальність дослідження зумовлена необхідністю розробки інтерпретабельних процес-орієнтованих моделей, які дозволяють оцінити зростання функції надійності залежно від розподілу зусиль. Мета роботи полягала в моделюванні процесів, пов'язаних із внесенням, виявленням та виправленням помилок засобами алгебри алгоритмів та нечіткої логіки. Запропонована методологія аналізу надійності програмного забезпечення базувалася на теорії надійності алгоритмічних процесів. Логіко-алгоритмічна модель процесу розробки побудована на основі лінійної, альтернативної та ітеративної операторних структур. Послідовність робіт без зворотного зв'язку описана лінійною структурою. Етапи верифікації та валідації описані за допомогою альтернативної та ітеративної алгоритмічних структур. Процес перевірки та виправлення, коли виявлені помилки негайно усувалися, а нові помилки не вносилися, описано альтернативною структурою. Процес налагодження, під час якого можуть вноситись нові помилки, описано ітеративною структурою. Логіко-алгоритмічна модель у вигляді нечіткої бази знань дозволила проектувати програмне забезпечення з необхідними рівнями надійності та витрат, використовуючи покращувальні перетворення. Система нечітких логічних рівнянь пов'язувала рівні правильності робочих, контрольних та доробочних операцій з можливістю правильного виконання процесу розробки. Розподіл зусиль формалізовано за допомогою покращувальних підстановок, введених у логіко-алгоритмічну модель. Керувальні змінні, пов'язані з покращувальними підстановками, інтерпретувалися як якість виконання робочих, контрольних та доробочних операцій. Запропонована нечітка модель надійності програмного забезпечення дозволила оцінити ризики процесу розробки на основі експертної та експериментальної інформації про надійність та часові характеристики етапів життєвого циклу. Нечітка модель була побудована шляхом перенесення надійних частин процесу розробки, отриманих з історій помилок та дефектів попередніх проєктів, у процес-орієнтовану модель надійності поточного проєкту. Розглянуто приклад аналізу надійності процесу розробки мобільного додатку для агрегації зображень, де впливовими факторами є безпомилкове виконання робочих, контрольних і доробочних операцій. Практичне значення дослідження полягає у створенні інструментарію, що дає змогу прогнозувати надійність програмного забезпечення на різних етапах його життєвого циклу, оптимізувати розподіл ресурсів між виявленням і виправленням помилок та зменшувати ризики невдалих рішень у проєктуванні і налагодженні

Ключові слова: надійність програмного забезпечення; оцінка ризиків процесу розробки; розподіл зусиль; логіко-алгоритмічна модель; нечітка модель надійності

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 15.08.2025 Revised: 11.11.2025 Accepted: 23.12.2025

UDC 004.056.5:004.032.26 DOI: 10.31649/vitce/3.2025.148

Synergy of artificial intelligence, SDN, Zero Trust, and blockchain: An overview of new trends in secure network management

Oleksandr Pidpalyi*

PhD
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
03056, 37 Beresteiskyi Ave., Kyiv, Ukraine
https://orcid.org/0009-0007-6852-7959

Oleksandr Romanov

Doctor of Technical Sciences, Lecturer
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
03056, 37 Beresteiskyi Ave., Kyiv, Ukraine
https://orcid.org/0000-0002-8683-3286

Abstract. The research relevance is determined by the need to create effective, transparent, and cyberattack-protected network management systems. The study aimed to systematise and critically analyse current approaches to combining artificial intelligence, software-defined networks, Zero-Trust architecture and blockchain to build adaptive, transparent and cyberattack-proof network management systems. A conceptual review of secure network management technologies was conducted using interpretative and comparative analysis of scientific sources, systemic and structural-categorical analysis of the characteristics of software-defined networks, Zero Trust architecture, blockchain, and artificial intelligence, and modelling scenarios for their application to improve the adaptability, transparency, and resilience of network systems in critical sectors of Ukraine. The results showed that the combined use of these technologies provides centralised traffic management, dynamic access policies, transparency of operations, and the ability to autonomously detect threats, significantly increasing the resilience of the network to multi-vector cyber-attacks. The study determined that the main problems of integrating these technologies into network systems are the opacity of artificial intelligence solutions, conflicts between the dynamism of models and the immutability of blockchain, high resource requirements, and the complexity of policy coordination in multi-domain networks. The implementation of Explainable Artificial Intelligence, hybrid architectures, off-chain solutions, model optimisation, and federated protocols has overcome limitations, providing a transparent, adaptive, and secure network system capable of responding effectively to threats and dynamic changes in the environment. The results showed that traditional solutions based on static firewalls and centralised control are limited in terms of response speed, attack detection accuracy and scalability. Integrated models combining artificial intelligence, software-defined networking, Zero-Trust architecture, and blockchain provide instant threat response, highly accurate attack detection, dynamic access control, automated auditing, and effective scalability, creating an adaptive, resilient, and transparent network system. The results of the study can be used to develop and optimise cybersecurity policies, automate access control and network event monitoring, and build scalable and transparent architectures of management systems

Keywords: network security; cyber defence; intrusion detection; machine learning; explainable artificial intelligence

Introduction

The growing complexity of cyber threats renders traditional network security models based on static policies and perimeter protection ineffective. Modern attacks bypass classic defence mechanisms, creating a need for dynamic solutions capable of detecting anomalies in real time and adaptively changing access policies. The combination of artificial intelligence (AI), Software-Defined Networking (SDN), Zero Trust Architecture (ZTA) and blockchain

Suggested Citation:
Pidpalyi, O., & Romanov, O. (2025). Synergy of artificial intelligence, SDN, Zero Trust, and blockchain: An overview of new trends in secure network management. *Information Technologies and Computer Engineering*, 22(3), 148-163. doi: 10.31649/vitce/3.2025.148

*Corresponding author

creates the basis for new approaches to security management. AI provides traffic and user behaviour analysis, SDN provides a flexible infrastructure, ZTA enables micro-segmentation and dynamic authentication, and blockchain ensures the immutability of event logs. For Ukraine, such synergy is necessary in the context of protecting critical infrastructure and government electronic services from hybrid threats.

The issue of integrating technologies to improve the adaptability and security of network systems was addressed by Ukrainian scientists. V.S. Nikitchenko (2024) studied the trends of digital transformation of business structures in the context of Industries 4.0 and 5.0, paying attention to the integration of modern technologies to improve the adaptability and efficiency of enterprises. The study demonstrated that the comprehensive implementation of automated control systems and decentralised technologies, such as blockchain and AI, can significantly increase the reliability of business processes, which directly correlates with approaches to the integration of SDN, ZTA and blockchain in secure network environments. These findings confirm the relevance of combining several technologies to ensure the adaptability and transparency of network systems. In turn, M.V. Vorokhob (2023) analysed models and methods for the improvement of enterprise security policies based on the Zero Trust methodology. The study emphasised that the implementation of behavioural access control models, continuous user monitoring and dynamic policy adaptation is key to reducing the risks of insider threats and ensuring the reliability of multi-domain networks. These concepts correlate with the use of AI+ZTA in practical scenarios of integrated network management, where AI behavioural algorithms assess user risk and automatically adapt access rights.

S.A. Latif *et al*. (2022) proposed a comprehensive security architecture for IoT networks of cyber-physical systems, integrating AI, blockchain, and SDN. The study demonstrated that this combination not only automates anomaly detection and attack prediction but also ensures real-time data transparency and immutability. These results confirm the feasibility of using multi-component solutions to improve the resilience of critical network infrastructures. At the same time, M.H. Bashaa *et al.* (2025) reviewed the integration of ZTA and machine learning (ML) to improve the security of software-defined networks. The study determined that the combined use of behavioural models, adaptive access policies, and AI for threat prediction significantly improves attack detection accuracy and enables rapid response to incidents. The authors also emphasise the relevance of audit and monitoring automation, which is consistent with practical cases of integrated network solutions implementation in various industries. L. Alevizos *et al.* (2022) considered the issue of extending ZTA to endpoints using blockchain. The study determined that the integration of decentralised registries increases the system's resistance to attacks and ensures audit transparency. The study emphasised that combining ZTA and blockchain helps protect critical network components and boosts trust in secure access management in corporate environments. This approach is relevant to the analysis of the potential of integrated network security solutions that combine ZTA and decentralised technologies.

S. Tiwari *et al.* (2022) proposed an approach to integrating AI with ZTA to improve the adaptability of network security in modern cyber threats. The study showed that AI can be used for real-time assessment of user and device behaviour and automatic adjustment of access policies, which significantly reduces the risk of security breaches. The study emphasised the potential of combining AI and ZTA to increase the flexibility and effectiveness of multi-domain network protection. B. Chowdhury *et al.* (2023) presented a conceptual model of a digital twin for e-Healthcare based on 6G using Zero Trust and blockchain. AI-driven attack prediction and response mechanisms, combined with decentralised registries, have been shown to enhance the security of critical healthcare systems. This approach demonstrates the practical benefits of integrated solutions in the context of protecting confidential data and ensuring service continuity. At the same time, A.V. Nagarjun & S. Rajkumar (2024) conducted a comprehensive review of the potential of deep learning and blockchain for intrusion detection systems (IDS). The study determined that the combination of AI and blockchain increases the accuracy and speed of anomaly detection while ensuring the transparency and immutability of event logs. This approach supports the idea of creating adaptive, autonomous security systems that can effectively respond to complex and multi-vector threats.

An analysis of previous studies demonstrated that existing studies are mostly limited to theoretical models or the analysis of individual corporate cases, without covering their interaction in scalable environments focused on the public sector and critical infrastructure. This creates a gap in the scientific and applied justification of integrated solutions that combine AI, SDN, Software-Defined Wide Area Network (SD-WAN), ZTA, and blockchain into a robust network management system. The study aimed to systematise and critically analyse current research on the integration of AI, SDN, ZTA and blockchain technologies to create adaptive, transparent and cyber-resilient network management systems. To achieve this goal, the following tasks were set: to identify and analyse existing approaches and models for integrating these technologies into network systems, to evaluate their effectiveness and limitations; to identify synergies between components, key challenges and prospects for the further development of integrated adaptive network solutions.

Materials and Methods

A conceptual review of modern technologies for secure network management and their integration was conducted to improve the adaptability, transparency, and resilience of telecommunications systems to multi-vector cyber threats. From the overall pool of scientific and scholarly

publications published between 2021 and 2025, a total of 34 academic sources met the inclusion criteria and were thematically relevant to the scope of the study. However, only a subset of these sources was directly employed in the formulation of comparative results and in the modelling of integrated network scenarios (AI + SDN + ZTA + blockchain). The remaining sources primarily served a supportive role, contributing to the development of the conceptual framework, the theoretical grounding of the study, and the enrichment of the discussion concerning limitations, risks, and future directions of technology integration. The literature search was conducted using major international scientific databases, including Scopus, Web of Science Core Collection, IEEE Xplore, ACM Digital Library, and ScienceDirect, which ensured comprehensive coverage of peer-reviewed research in the fields of network engineering, cybersecurity, and information systems.

Criteria for inclusion of sources works describing the integration of these technologies into network systems, research on Zero Trust architectural solutions, the use of AI for threat prediction or security policy automation, as well as reviews and comparative studies of security models. Criteria for excluding sources: publications that do not contain specific data on the integration of technologies or their impact on the adaptability and resilience of networks, works that deal exclusively with hardware solutions without elements of SDN, ZTA, blockchain or AI, and materials published before 2021. The research was conducted from March to August 2025.

The research methodology involved systematising and conducting a comparative analysis of the characteristics of each technology. The method of interpretative analysis of scientific sources was used to evaluate the architecture, tasks and functional capabilities of AI, SDN, ZTA and blockchain, as well as the method of comparative analysis to compare their advantages and limitations in the context of building integrated network solutions. To structure the data obtained, a method of systematic and structural-categorical analysis was used, which facilitated the organisation of technology characteristics into logical blocks and the creation of analytical tables. This facilitated a detailed description of the key functions of the technologies, their application scenarios, advantages for security and network management, potential limitations and ways to overcome them, as well as a comparison of traditional and integrated security models.

Modelling of integrated scenarios for the application of technologies in networks was highlighted. To assess the advantages of integrated solutions, an analysis of technical and organisational aspects was conducted, including increased network adaptability and flexibility, automated threat detection and access control, ensuring data transparency and immutability, as well as scalability and integration into multi-domain networks. Potential challenges and limitations of technology integration were also explored, including explainable AI issues, conflicts between AI and blockchain dynamics, high computing resource requirements and policy coordination in multi-domain networks, as well as ways to overcome them. A comparative analysis of traditional network security models and integrated solutions (AI + SDN + ZTA + blockchain) was conducted based on the criteria of incident response speed, attack detection accuracy, scalability, resistance to internal threats, transparency and auditability, adaptability to dynamic changes, and level of automation.

The study examined application-orientated conceptual scenarios for deploying integrated AI, SDN, ZTA, and blockchain in Ukraine. These scenarios were not treated as fully documented, organisation-specific case studies with proprietary network datasets; rather, they represent desk-based modelling and feasibility assessment grounded in (i) the reviewed scientific literature on AI/SDN/ZTA/blockchain integration and (ii) open policy and industry documents that describe reference architectures, maturity targets, and automation principles for multi-domain networks. Three scenario classes were analysed. First, public e-service delivery environments (including high-assurance digital service platforms) were modelled as ZTA-enabled service perimeters in which AI supports anomaly detection and risk-based access decisions, while a permissioned ledger provides tamper-evident audit trails for security-relevant events. Second, critical infrastructure communications and control-support networks were analysed at the level of architectural patterns: SDN enables rapid traffic engineering and segmentation, AI performs predictive detection of abnormal behaviour, and ledger-based logging strengthens traceability and accountability of configuration and access actions. Third, multi-domain government networks were considered as simulated, federated environments within a conceptual modelling framework, in which ZTA enforces continuous verification, AI automates behavioural monitoring, and blockchain-backed audit logs enhance cross-domain accountability. Consequently, the information basis for these scenarios was derived primarily from analysed reports, standards, and industry architecture documents, while the academic corpus was used to substantiate technical feasibility, integration constraints, and expected effects. These cases are therefore reported as simulation-based conceptual implementations rather than empirical evaluations of named Ukrainian networks or operational systems.

The study also developed recommendations for the development of a national AI ecosystem, standardisation and integration of technologies, staff training and cooperation with international partners to implement integrated solutions in critical infrastructures. In addition, industry reports and documents regulating the implementation of integrated network solutions and approaches were analysed. The Memorandum for the Heads of Executive Departments and Agencies (2022), DoD Zero Trust Strategy (2022), AT&T Domain 2.0 Vision White Paper (2013), and Telefónica (2017) approaches were reviewed. In addition, LF Networking projects focused on the Open Network Automation Platform (Alhilali & Montazerolghaem, 2023) were analysed. The analysis of these cases revealed real

models of AI, SDN, ZTA, and blockchain integration, security policy standardisation, automation and audit principles in multi-domain networks, and key practical approaches to improving the cyber resilience and adaptability of network systems. Visualisation and modelling of data flows in the network were conducted using block diagrams illustrating the sequence of traffic processing, cyclical interaction of components, feedback mechanisms, and real-time self-regulation of the system. This was to assess not only the functional capabilities of individual technologies, but also their synergy in ensuring adaptive and autonomous network management.

Results and Discussion

Overview of basic technologies for secure network management

The review demonstrated how key components such as SDN, ZTA, blockchain, and AI interact to enable adaptive and secure network management. These technologies complement each other, providing centralised management, dynamic access policies, operational transparency, and autonomous threat detection capabilities. SDN represents an architecture with a separation of control plane and data plane; the SDN controller centrally manages traffic flows and sets routing policies on switching devices. AI integration can detect traffic anomalies, predict possible attacks, and dynamically redirect flows based on network conditions. SDN also provides event logging and real-time monitoring of network resources. ZTA implements the "zero trust" principle: the network is segmented into isolated zones (microsegmentation), and each user and device undergoes adaptive authorisation with constant verification. Access control is based on behavioural models and risk-oriented algorithms, which can be used to change access rights quickly depending on the threat (Aramide, 2024). For a systematic comparison of key secure network management technologies, their main characteristics, functions and integration capabilities are summarised in Table 1.

Table 1. Key features of technologies for secure network management

Technology	Key functions	Integration with other components	Main benefits	Potential limitations
SDN	Centralised traffic management, dynamic routing	AI for anomaly prediction and policy optimisation	Adaptability, quick response	Dependence on the controller, high resource requirements
ZTA	Adaptive access control, micro-segmentation	AI for behavioural assessment of users	Reduction of internal threats, dynamic authorisation	Complexity of implementation, need for constant verification
Blockchain	Decentralised storage, smart contracts	AI for log verification and auditing	Transparency, data consistency	Conflict with AI dynamics, transaction delays
AI (ML/DL/RL)	Traffic classification, anomaly detection	SDN for routing, ZTA for access, blockchain for logs	Automation, adaptability, threat prediction	Explainability (XAI), need for computational resources

Note: DL – deep learning; RL – reinforcement learning
Source: compiled by the authors based on analysis of data of H. Han *et al*. (2021), P. Svensberg (2023), A. Alshehri *et al*. (2024), F. Ashfaq *et al.* (2025)

Analysis of Table 1 shows that each considered technology is specific but complementary in secure network management. SDN provides centralised traffic management and provides a rapid response to changes in the network, but its dependence on a controller and high computing resource requirements are potential limitations. ZTA effectively reduces the risks of internal threats through adaptive access control and micro-segmentation, but it requires constant user verification and complex policy configuration. Blockchain increases data transparency and immutability and provides automated auditing, but the dynamic nature of AI models can conflict with transaction delays and the need for consensus in the network. AI in network solutions provides threat prediction, traffic classification, and adaptive security policy management, but requires explainability of decisions and significant computing resources.

The interconnection of these technologies compensates for the limitations of individual components: SDN integration with AI provides dynamic routing and anomaly detection; the combination of AI and ZTA ensures adaptive access control; blockchain supports transparency and immutability of actions. Thus, the combined use of SDN, ZTA, blockchain, and AI forms a comprehensive system that simultaneously increases the security, adaptability, and reliability of the network infrastructure. O. Aramide (2022) examines the principles of Zero Trust identity with continuous AI verification in next-generation networks. The study specifies that AI integration enables the creation of secure digital ecosystems with adaptive access control based on behavioural identities. This directly correlates with the ability to create secure digital ecosystems with adaptive access control based on behavioural identities. The study specifies that AI integration can be used to create secure digital ecosystems with adaptive access control based on behavioural identities. This directly correlates with the current approach to integrating XAI models into access policies: the study also emphasises the role of explainability and continuous verification. At the same time, O. Aramide focuses primarily on the user identity level, while the current study pays considerable attention to the network plane (SDN, policy-as-code). Figure 1 shows an integrated network architecture that combines SDN, the ZTA concept, and blockchain technologies.

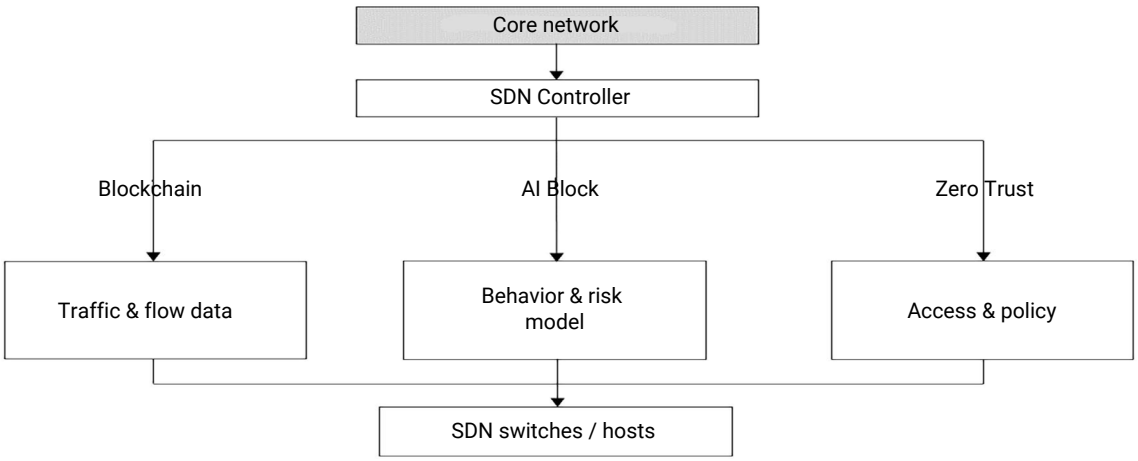


Figure 1. AI block integration diagram
Source: compiled by the authors based on analysis of F. Ashfaq *et al.* (2025)

The AI block is located between the SDN controller and the Zero Trust and blockchain system components. It performs adaptive analysis of traffic and user behaviour, which can be used for dynamic changes to routing and access policies. Log and transaction data are stored in blockchain, and AI analyses it for anomalies and threats. Interaction with the Zero Trust module ensures continuous verification of users and devices, while SDN provides flexible flow management. This integration creates a closed loop of adaptive network management with a high level of transparency, automation, and resistance to cyberattacks. To illustrate the interaction of key technologies in secure network management, a diagram is provided that shows the data flows and roles of each component (Fig. 2).

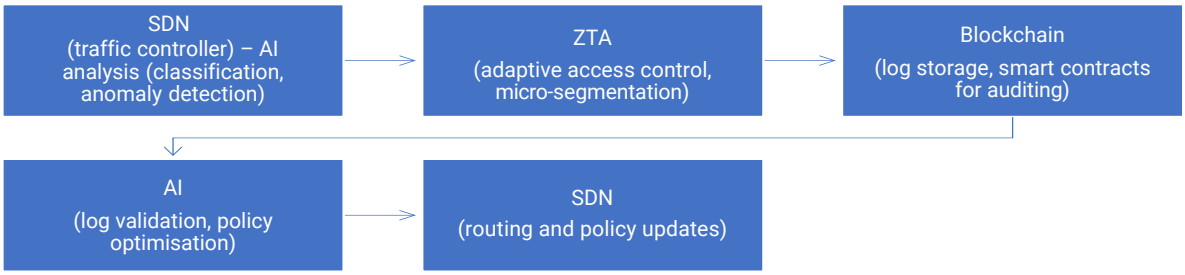


Figure 2. Integration scheme for technologies in a secure network
Source: compiled by the authors based on analysis of P. Svensberg (2023)

An analysis of the SDN, ZTA, blockchain, and AI integration scheme demonstrates a clear sequence of component interactions and the synergistic operation of all elements of the secure network management system. The SDN controller centrally manages traffic flows and provides primary data routing, while the received packets are sent for processing by AI models for traffic classification, anomaly detection, and potential threat prediction. Based on the AI results, the ZTA system adaptively adjusts user and device access rights using micro-segmentation and dynamic authorisation policies. All transactions and actions are recorded in the blockchain, which ensures data immutability, transparency, and the ability to perform automated audits through smart contracts. AI also interacts with blockchain to verify logs, evaluate policy correctness, and correct SDN routing in real time. This cyclical process creates a dynamic, adaptive, and attack-resistant network system where each technology compensates for the limitations of the others. SDN provides centralised management, ZTA enhances access security, blockchain ensures data transparency and immutability, and AI coordinates adaptability and threat prediction. As a result, the integrated system can respond to multi-vector cyberattacks, dynamically change routes and access policies, and maintain a high level of trust in network interactions.

Conceptual overview of scenarios for the application of integrated network technologies

Modern telecommunications systems face the need to ensure high adaptability and resistance to multi-vector attacks. The integration of AI, SDN, ZTA, and blockchain technologies creates new opportunities for automating network management, dynamic access control, and ensuring transparency of operations. To systematise these approaches, basic application scenarios were developed to assess the role of each technology and the effectiveness of their synergy in various aspects of security and network management (Table 2).

Table 2. Conceptual overview of scenarios, technologies, functions and advantages of using integrated network technologies

Use scenario	Employed technologies	Primary functions	Security and network management benefits
Dynamic routing	AI + SDN	ML models analyse traffic and adjust routing rules via an SDN controller	Optimisation of data flows, rapid response to anomalies, and increased network bandwidth
Adaptive access control	AI + ZTA	Behavioural models assess user and device risk, dynamically changing access rights	Reduction of internal risks, micro-segmentation, and increased trust in access
Audit and logging	AI + blockchain	Logging of all actions in a decentralised registry, analysis of logs for anomalies	Data integrity, transparency of operations, and rapid detection of incidents
Comprehensive scenario	AI + SDN + ZTA + blockchain	Integration of all components: threat detection, access policy adaptation, logging	Autonomy, resistance to multi-vector attacks, real-time adaptability, transparency and auditing

Source: compiled by the authors based on O. Aramide (2022) and S. Narayanan (2025)

The table shows four key scenarios for technology integration. In the dynamic routing scenario, ML models act as the analytical core, evaluating network flows and determining optimal routes. SDN is central in quick application of these decisions to network controllers, reducing latency and avoiding congestion. The adaptive access control scenario shows how AI and ZTA collaborate to evaluate user and device behaviour patterns, identifying risks in real time. This enables dynamic access rights management and micro-segmentation, which is critical for protecting internal network segments from potential threats. The audit and logging scenario demonstrates the advantages of blockchain combined with AI: immutable records and smart contracts ensure transaction transparency, while AI analyses logs to quickly detect anomalies and potential incidents. A comprehensive integration scenario for all technologies ensures maximum synergy: SDN centrally manages traffic flows, AI predicts threats and optimises policies, ZTA adaptively controls user access, and blockchain provides reliable auditing and immutability of logs. This approach creates an autonomous, resilient, and transparent network system capable of responding to changes in user behaviour, network load, and new types of threats in real time. A general analysis of the table shows that each technology performs a specific but interrelated function, and their synergy improves security and network management efficiency compared to traditional static solutions. The complexity of implementing such scenarios requires careful balancing of resources, policy coordination, and explainable AI to increase trust in decisions (Chaudhry, 2025). At the same time, S. Batewela *et al.* (2025) examined the challenges of security orchestration in next-generation networks. They conducted a comprehensive review of existing approaches and emphasised the need for integrated solutions capable of coordinating security policies across different network domains. This correlates with the current SDN-based "policy-as-code" approach and closed loops with AI; the contribution of the conducted research is to add immutable auditing and cross-agency interoperability through federated protocols. While scientists address operators and SOAR/IBN chains, the presented analysis details how to combine these chains with blockchain without losing response time (off-chain + periodic commit to the registry).

To illustrate the interaction of technologies in network systems, a block diagram was created that shows data flows, the roles of each component, and cyclical interaction in real time (Fig. 3). The diagram illustrates how AI analyses traffic and predicts threats, SDN provides centralised route management, ZTA implements dynamic access control, and blockchain ensures transparency and immutability of logs. It can be used to evaluate both the sequence of data processing and parallel flows, emphasising the complexity of integrated solutions.

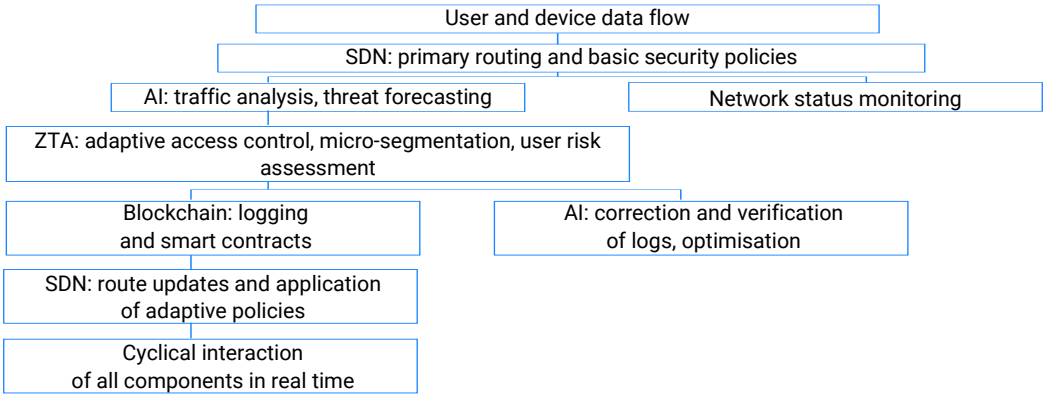


Figure 3. Interaction of technologies in network systems

Source: compiled by the authors based on S. Batewela *et al.* (2025)

The figure shows the step-by-step processing of network flows and the interaction of technologies. At the first level, user and device data are sent to SDN, which performs initial routing and applies basic security policies. At the same time, network status monitoring and traffic analysis are performed through AI modules, which can detect anomalies and predict potential threats. At the second level, ZTA implements adaptive access control, assessing user and device risks in real time, while blockchain records all events in a decentralised registry, ensuring the immutability and transparency of records. AI is used to verify blockchain logs and correct access and routing policies, creating a feedback and self-regulating mechanism for the system. At the third level, SDN applies adaptive routing policies, and redirects flows, and the interaction cycle is repeated in real time, ensuring constant adaptation of the system to changes in user behaviour, network topology and potential threats. Analysis of the diagram shows that each technology performs a specific but interrelated function: SDN is responsible for operational flow management, AI for analysis and forecasting, ZTA for adaptive access control, and blockchain for auditing and transparency. This architecture ensures system autonomy, high adaptability to changes in the network environment, and comprehensive protection against multi-vector attacks, while emphasising the need for policy coordination and optimisation of computing resources.

Analysis of the advantages of integrated solutions in network security

Integrated network security solutions enable a comprehensive approach to protecting digital infrastructures by combining different technologies into a single flexible system. This approach not only provides resilience against the growing number of cyber threats but also improves resource management efficiency and network scalability. First, integrated systems significantly increase network adaptability and flexibility by enabling rapid response to new threats, real-time changes to access and protection policies, and adaptation to different environments. This is a priority for dynamic infrastructures such as cloud services or corporate multi-domain networks. The second advantage is the automation of threat detection and access management, which minimises the human factor and reduces response time. The use of AI and ML can quickly identify traffic anomalies, block malicious actions, and promptly update security rules. Transparency and immutability of data are crucial, which is achieved using blockchain and distributed ledger technologies. This ensures trust in audit results, prevents unauthorised interference, and creates conditions for the formation of a unified information picture across the entire organisation. Lastly, integrated solutions provide scalability and integration into multi-domain networks where different technologies and protocols are used simultaneously. Thanks to their modularity and flexible architectural approaches, such systems can be easily expanded without losing performance and functionality. This creates the basis for the sustainable development of digital infrastructures in the future. For clarity, the main advantages of integrated solutions in network security are presented in Table 3.

Table 3. Advantages of integrated network security solutions

Area of expertise	Key aspects	Importance of network security
Improving network adaptability and flexibility	Dynamic routing, load balancing, integration with SDN	Ensures rapid response to traffic changes, reduces the risk of overload and downtime
Automated threat detection and access control	Using AI/ML for IDS/IPS, Zero Trust, and automated access policies	Minimises human error, accelerates attack neutralisation, and guarantees access control at all levels
Ensuring transparency and consistency of data	Real-time logging and monitoring, blockchain technologies, and SIEM platforms	Increases trust in the system, prevents unauthorised changes, and ensures the verifiability of events
Scalability and integration into multi-domain networks	Cloud and hybrid environments, modular architectures, API integrations	Ensures flexible infrastructure expansion and simplifies management of complex distributed networks

Source: compiled by the authors based on R. Dwivedi *et al*. (2023), A. Malik *et al*. (2025)

The table shows that the advantages of integrated solutions cover both technical and organisational aspects of security. The combination of adaptability and flexibility can ensure network resilience even in rapidly changing cyber threat environments. Automation significantly reduces response time to attacks and reduces dependence on the human factor. Transparency and data immutability increase the level of trust in the system from both users and regulators, which is a priority in the context of regulatory compliance. Lastly, scalability and multi-domain integration make the system flexible from a strategic perspective, which can be used for quick expansion without significant investment in infrastructure restructuring. Thus, integrated network security solutions are not only a technological tool, but also a strategic approach to building secure and flexible digital environments.

Assessment of challenges and limitations of network technology integration

The integration of AI, SDN, ZTA, and blockchain technologies comes with a bunch of challenges, both technical and organisational. One key thing is XAI, as modern ML and DL models often work like “black boxes”, which restricts the ability of administrators and analysts to determine the logic behind the decisions. The opacity of AI algorithms complicates auditing, control, and user trust, which can

potentially lead to incorrect or untimely responses to threats. The use of XAI, standardised documentation of decisions and visualisation of model logic ensures transparency, controllability and soundness of decision-making, which in turn contributes to increased integration efficiency and trust in the system on the part of users and administrators. There is a significant conflict between the dynamism of AI and the immutability principle of blockchain technologies. AI models require constant updating for adaptive network policy management and rapid response to threats, while blockchain ensures the immutability of records and transactions. This creates a potential contradiction that can hinder the synchronisation and coordination of network processes. To overcome this, hybrid architectures, off-chain update mechanisms, and conditionally adaptive smart contracts are used to maintain AI adaptability while ensuring the immutability of data and logs (Speith, 2022).

Another substantial limitation is the high resource intensity of integrated solutions, which arises from the simultaneous use of complex AI models and decentralised blockchain registries. Increased load on computing resources can reduce system performance, increase data processing delays, and limit network scalability. To optimise these processes, lightweight model architectures, pruning and quantisation, as well as distributed computing and cloud computing platforms, are used. This approach ensures effective network adaptability while maintaining response speed and threat prediction accuracy. In addition, the coordination of security policies in multi-domain networks remains critical, as different domains or organisations may use different access, control, and authentication standards. The lack of uniform protocols can lead to conflicts, duplication of rules, and reduced effectiveness of integrated solutions. The use of federated protocols, unified standards, and integration mechanisms can coordinate policies across domains, maintain centralised control, and preserve the autonomy of individual network segments (Pemmasani *et al.*, 2025). This provides a balance between flexibility, security, and compatibility of heterogeneous network environments. Below is a summary Table 4, which systematises the key challenges and possible ways to overcome them.

Table 4. Challenges, problems, limitations and ways to overcome them

Challenge/limitation	Issue	Potential solutions
XAI	AI algorithms are often "black boxes"; it is difficult to explain their decisions	Implementation of XAI, decision documentation standards, and model visualisation
Conflicts between the dynamism of AI and blockchain	AI requires model updates, and blockchain ensures immutability	Hybrid architectures, off-chain solutions, conditionally adaptive smart contracts
High demands on computing resources	AI+blockchain requires significant resources	Cloud services, distributed computing, model optimisation
Policy coordination in multi-domain networks	Different domains have unique security standards	Federated protocols, integration standards, unified access rules

Source: compiled by the authors based on Z. Azam *et al.* (2023)

Following the table, these challenges are closely interrelated: the transparency of AI decisions affects integration with blockchain, resource constraints determine the need for optimisation, and the use of distributed computing and multi-domain conflicts requires the unification of standards and protocols. The comprehensive use of the proposed solutions creates an adaptive, transparent, and secure network system capable of effectively responding to modern threats and dynamically changing operating conditions.

Comparative analysis of integrated and traditional security models

Traditional network security models are based on static mechanisms such as firewalls, IDS, and centralised access control systems. They apply rigidly defined rules and policies, which limit their ability to adapt to dynamic environments and multi-vector cyber-attacks. Static firewalls are efficient against known threats, but they are unable to respond quickly to new types of attacks or internal incidents. Centralised solutions control resources from a single location, but in large and distributed networks, response speed and performance are significantly reduced.

Integrated security models that combine AI, SDN, ZTA, and blockchain offer a more comprehensive approach. SDN provides centralised traffic flow management and dynamic routing, enabling rapid response to network changes. AI modules automatically analyse traffic, classify packets, detect anomalies, and predict potential threats in real time. ZTA provides adaptive access control and micro-segmentation, reducing the risks of internal threats, while blockchain ensures transparency and immutability of records, automating the audit and verification of user and device actions (Hashmi *et al.*, 2025).

A comparison of these two approaches reveals fundamental differences (Table 5). In traditional systems, incident response speed is limited by manual intervention, attack detection accuracy depends on predefined rules, and scalability and resilience to internal threats are significantly limited. Integrated models provide instant response thanks to AI and SDN, enable proactive detection of new and complex attacks, scale easily in multi-domain and cloud environments, and ensure comprehensive protection against internal and external threats through continuous access control and transparent auditing.

Table 5. Comparative analysis of traditional and integrated network security models

Parameter	Traditional models (static firewalls, centralised solutions)	Integrated models (AI + SDN + ZTA + blockchain)
Incident response speed	The response to events takes from a few minutes to hours; automation is limited to simple rules. For example, during a DDoS attack, manual traffic redirection	Reaction within a second thanks to AI that predicts attacks and SDN that dynamically redirects traffic. For example, AI detects traffic anomalies, and SDN changes routes to reduce load
Accuracy of attack detection	60-70% detection of known attacks; new threats are missed due to static signatures	90-95% thanks to the combined use of ML/DL for traffic analysis, ZTA behavioural patterns and log verification in blockchain
Scalability	Limited by centralised controllers, it is difficult to maintain multi-domain networks. Additional equipment and manual configuration are required for network expansion	High; SDN enables centralised management of thousands of switches, AI automatically adapts policies, and blockchain ensures log consistency across multi-domain systems
Resilience to internal threats	Low; control is limited by ACL rules or basic authorisations	High; ZTA continuously verifies users and devices, AI assesses risks in real time, and blockchain stores immutable records of all events
Transparency and audit	Limited by centralised logs, data modifications are possible in the event of server compromise	Complete transparency thanks to blockchain; all transactions are recorded, smart contracts automatically verify actions, and AI analyses logs for anomalies
Adaptability to dynamic changes	Non-existent; changes in topology, load or new threats require manual intervention	High: AI predicts traffic and threats, SDN dynamically changes routes, and ZTA adapts access rights in real time
Level of automation	Low; constant intervention by administrators is required to change rules, monitor and audit	Maximum; AI manages traffic classification, threat prediction, access policy adaptation, and blockchain provides automatic auditing without human intervention

Note: ACL – Access Control List
Source: compiled by the authors

The table shows that integrated security models significantly outperform traditional solutions in all key criteria. The advantage of integrated systems includes not only faster response to incidents and high accuracy in detecting attacks, but also the ability to perform automated audits, dynamic access control, and network infrastructure scaling. SDN enables centralised and efficient traffic flow management, AI predicts threats and adapts security policies, ZTA provides continuous user and device verification, and blockchain ensures transparency and immutability of records. Thanks to the synergy of these technologies, integrated models create an adaptive, resilient, and transparent network system capable of responding quickly to multi-vector threats and providing reliable protection for internal and external resources. Compared to classic static solutions, such integration can optimise resources, reduce response times, and increase the level of trust that users and administrators have in network security. At the same time, analysis of current publications confirms a common trend: a shift from static perimeter approaches to integrated architectures, where AI is responsible for threat analysis and prediction, SDN for dynamic traffic orchestration, ZTA for continuous access verification, and blockchain for transparent and immutable auditing. The results obtained are consistent with this vector and further emphasise the practical importance of explainable models, hybrid (on-/off-chain) logging schemes, and policy-as-code for closed-loop real-time control.

Conceptual overview of potential implementation cases for Ukraine and recommendations for optimising integrated network solutions

Ukraine has already formalised key elements of information security and digital trust management systems at the state sector level (Information Security Management System, qualified electronic trust services, centralised identification tools) and has direct experience in countering coordinated cyber operations against energy and telecommunications. Therefore, the next stage is the transition from perimeter-based, predominantly static models to "continuous verification" modes and policies that are data-driven and automatically applied through an SDN network factory in real time. The basis for such a transition is provided, on the one hand, by proven certification of IB processes in state organisations and services, and on the other hand, by mature open standards for state Zero-Trust transformations, which record control states of maturity by domains of identities, devices, networks, applications and data. Below is a conceptual overview of the possible application of these technologies in critical sectors of Ukraine (Table 6).

Table 6. Potential cases for Ukraine

Potential case	Technological components (AI, SDN, ZTA, Blockchain)	Expected result
Adaptive management of public electronic services	AI (anomaly detection, access control), integration with blockchain for data protection	Improved cyber resilience, minimised fraud, optimised service delivery processes
Protection of energy and telecommunications infrastructure	SDN (dynamic traffic management), AI (attack prediction), blockchain (transaction transparency, data protection)	Continuity of critical systems, reduction of cyberattack risks to power grids and mobile networks
Implementation of ZTA in multi-domain networks of state authorities	ZTA (Zero Trust), AI (automated user behaviour monitoring), blockchain (access auditing)	Reducing insider threat risks, securing access to resources, and controlling interdepartmental exchanges

Source: compiled by the authors

The first direction is aimed at modernising the portfolio of public e-services by integrating ZTA as an operational access "skeleton", SDN as a micro-segmentation and routing network factory, and AI as an analytical core for risk assessment and behavioural validation of requests. The organisational prerequisite is already in place: most state-owned enterprises were certified under the Information Security Management System according to ISO/IEC No. 27001 (2022), which formalises the Plan-Do-Check-Act (PDCA) cycle and standardises risk, incident and change management artefacts. In addition, the state segment of trust services is supported by the Central Certification Authority and qualified signature services, which simplifies the unification of trust roots and interagency interaction policies. On the technological level, this technology can be used to build a Zero-Trust gateway as a control plane that aggregates signals from proxies, identity gateways, and API brokers, correlates them in ML models, and then transmits them to the SDN controller for microsegmentation, Quality of Service (QoS) and routing policies based on subject context, resource sensitivity and current risk. Within the adjunct security bus, key events such as privilege escalations, behavioural profile deviations, and access policy changes are logged in a permissioned registry with minimal impact on transaction latency, creating a reproducible trail for compliance auditing and forensics. External benchmarks, from the Memorandum for the Heads of Executive Departments and Agencies (2022) to the DoD Zero Trust Strategy (2022), set specific control targets that can be used to build a roadmap for the maturity of the Ukrainian GovTech segment without the need to replicate already known approaches. As a result, this architecture puts the "user-service-data" interaction into a mode of constant verification and provides the basis for the "AI for Access Governance" pilot, in which XAI deterministically justify both an increase in the level of trust and the imposition of additional authentication factors.

The second case concerns energy and telecommunications and responds to the specifics of hybrid threats that combine targeted attacks on ICS segments with the destruction of public networks. Incidents involving the Industroyer/Industroyer2 family targeted energy network automation and demonstrated the ability to disrupt technological protocols, while in December 2023, a cyberattack on the largest mobile operator – Kyivstar – led to large-scale service disruptions comparable to "national-level failures" and demonstrated the need for policy-driven segmentation, domain isolation, and rapid recovery of controllability (Chuzavkov, 2023). The architectural response is that SDN deploys controllable overlays between the technology and business zones, while AI closes the feedback loop: it correlates telemetry from the core, transport and Radio Access Network segments, generates policy-as-code decisions and initiates the reconfiguration of routes, ACLs and QoS via the controller. For telecom operators, industry-proven virtualisation and automation programmes such as AT&T Domain 2.0 Vision White Paper (2013), Telefónica (2017), and LF Networking projects already describe target models with "closed control loops" where policies are formed based on event flow and applied automatically. For the energy sector, a relevant component is the constant auditing of commands and configuration changes through a permissioned registry. This approach has been verified in European pilots by Energy Web/TenneT to increase observability and reduce the cost of regulatory investigations. Together, this integration reduces detection and localisation time, increases the evidential value of operator and process attribution, and provides a "controlled degradation" mode during large-scale incidents through rapid isolation/segmentation scenarios.

The third vector is multi-domain networks of government agencies, where ZTA is the operational "constitution" of access, while SDN and AI provide the engine for real-time segmentation, orchestration, and policy adaptation. In practice, this implies the unification of trust roots, the federation of identity attributes between central and local authorities, the introduction of end-to-end visibility of access transactions, and the creation of service catalogues with clear separation of responsibilities in object-level access policies. For accelerated implementation, it is advisable to deploy in parallel an interagency trust mesh for transactional data exchange and an SD-WAN overlay for geographically dispersed institutions, where policies are propagated through ONAP-compatible interfaces and applied in the form of code, and all changes, from delegation of rights to temporary exceptions, are recorded in an immutable registry to simplify auditing and forensics. Methodological guidelines for government implementations are already available in the form of public Zero Trust strategies and analytics on cyber operations against

Ukraine, which emphasise the need to move from static standards to context-adaptive control modes. This can be used to form a step-by-step PDCA plan with milestones, metrics, and maturity profiles at the level of identity domains, devices, networks, and data. Combined with institutional support and professional clusters, this creates new opportunities for standardised deployments in the state's production circuits. A. Gupta *et al.* (2023) proposed the use of proxy smart contracts to implement Zero Trust principles in decentralised oracle networks. They explored how this approach enables secure data exchange between decentralised applications, minimising the risks of data manipulation or falsification. This correlates with the current conclusion about the role of blockchain as an integral layer of trust and audit, but the analysis by A. Gupta *et al.* of smart contract templates in the specific domain of DON does not cover the SDN network plane and micro-segmentation policies that were key in the model under consideration. The discrepancy is due to different system granularity.

In summary, these trajectories demonstrate that the integration of AI, SDN, and ZTA, with the measured use of blockchain as an immutable audit mechanism, forms an operational architecture capable of simultaneously accelerating detection and response, improving detection and attribution accuracy, scalable microsegmentation of service domains, and procedural transparency in security decision-making. The Ukrainian context, from proven certifications in the public sector and the compatibility of trust services to lessons learned from real cyberattacks against the energy and telecommunications sectors, provides both empirical grounds and practical markers of maturity for moving from pilots to systemic implementations, using open standards and industry roadmaps for network virtualisation and automation.

Optimisation of integrated network solutions requires a systematic approach that combines the development of technological infrastructure, human resources and international cooperation. The development of a national AI ecosystem is critical, as the implementation of AI Factory will enable the creation of autonomous cyber defence systems capable of responding quickly to new threats. The standardisation and integration of SDN, ZTA and blockchain technologies must be based on harmonisation with Ukrainian legislation and incorporate international compatibility and security requirements. Investing in training specialists is a priority, as the lack of qualified personnel is one of the main barriers to the development of comprehensive cybersecurity solutions. Lastly, international cooperation accelerates the implementation of new technologies and avoid duplication of efforts by using best practices and standards that have already been tested by other countries. To summarise the strategic directions for the development of the national cyber defence system, they have been systematised, correlating each benchmark with specific implementation mechanisms and expected results (Table 7).

Table 7. Recommendations and directions for their implementation

Direction	Specific implementation	Expected result
Development of the national AI ecosystem	Development of AI Factory, investments in data centres, creation of platforms for testing AI solutions	Autonomy, reduced dependence on external providers, and increased cyber resilience
Standardisation and integration of technologies	Development of national standards for SDN, ZTA, and blockchain integration; harmonisation with international standards	Ensuring compatibility, data protection, and adaptation to Ukrainian realities
Training and professional development of personnel	Certification programmes, creation of educational courses, partnerships with universities and international organisations	Formation of a highly qualified personnel reserve, reduction of personnel shortages
Cooperation with international partners	Participation in joint research, exchange of technologies, and creation of consortia in the field of cybersecurity	Acceleration of innovation, access to advanced technologies, and increased global integration

Source: compiled by the authors based on A. Kudriashov (2024)

Analysis of the table shows that the development of a national AI ecosystem through the creation of an AI Factory and corresponding data centres ensures autonomy and reduces dependence on external technology providers, while increasing the speed of anomaly detection and response to cyber threats. S. Mishra (2023) proposed a hybrid IDS system based on blockchain and ML to protect "smart" networks and maintain confidentiality. The system analyses traffic in real time, detects anomalies, and stores logs in a blockchain for transparent auditing. The results showed that the combination of ML and blockchain significantly improves attack detection efficiency while maintaining user data privacy. This is consistent with present findings on the feasibility of hybrid on-/off-chain approaches and AI analytics to ensure trust and reduce attack risks. The difference is in the level of application: in S. Mishra, IDS is the primary security tool, while in the present model, it is integrated into a broader Zero Trust perimeter alongside SDN and XAI. The standardisation and integration of SDN, ZTA and blockchain technologies ensures compliance with international standards and Ukrainian legislation, which can be used for the construction of a secure and flexible network architecture with minimal conflicts between different systems. Investments in training and certification of personnel create a highly qualified personnel reserve capable of effectively managing complex integrated systems, which is critical for

the stability of the national infrastructure. Active international cooperation accelerates innovation, the exchange of best practices, and access to advanced technologies, increasing the global integration and adaptability of Ukraine's cyber system. Overall, the table demonstrates a comprehensive approach in which technological, organisational, and human resources components are interrelated and mutually reinforce the effect of increasing cyber resilience.

A comprehensive approach to optimising integrated network solutions, where technological, organisational and human resources components are interconnected to enhance cyber resilience, is presented in Figure 4. The diagram illustrates the interaction between the development of the national AI ecosystem, the standardisation and integration of SDN, ZTA and blockchain technologies, investment in specialist training and international cooperation. It demonstrates how these areas support each other, creating an adaptive, scalable and secure network infrastructure capable of responding effectively to modern cyber threats and ensuring the resilience of state systems.

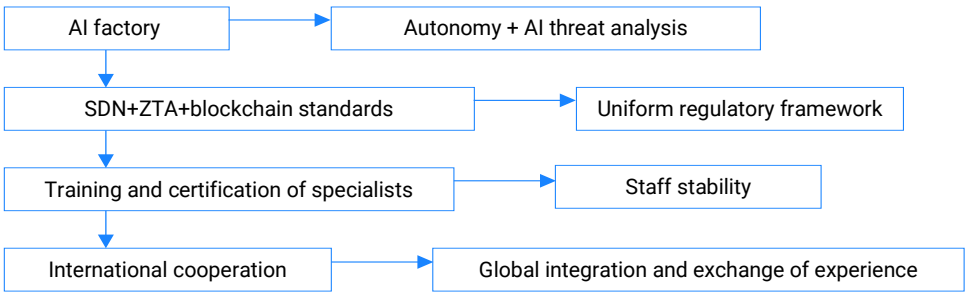


Figure 4. Optimisation of integrated network solutions
Source: compiled by the authors based on the data of H. Li *et al*. (2025)

Analysis of the diagram shows that successful optimisation of integrated network solutions is based on the synergy of technological, organisational and human resources components. The development of a national AI ecosystem ensures autonomy and rapid detection of anomalies, while the standardisation and integration of SDN, ZTA and blockchain guarantee the compatibility and security of the network infrastructure. Investments in staff training increase the level of expertise and readiness to respond to cyber incidents. International cooperation accelerates the implementation of advanced technologies and can be used to adapt best practices to the Ukrainian context. The diagram shows that the comprehensive combination of these elements creates an adaptive, scalable, and threat-resistant network architecture, where each component enhances the effectiveness of the others, providing a comprehensive system of state-level cyber protection.

In the doctoral thesis, C. Katsis (2025) developed a comprehensive framework for specifying, training, and implementing network-wide access control in Zero-Trust Network Architectures. The study described a methodology for building access policies, model training algorithms for dynamic control, and automated security enforcement mechanisms. The research confirmed that centralised policy specification combined with automated learning can improve scalability and security effectiveness in complex corporate networks. This is directly consistent with the current approach to policy-as-code and AI integration for automatic real-time rule updates. The difference is that C. Katsis emphasises the formalisation and modelling of policies, while the study also focuses on XAI and blockchain integration for auditing. The difference is due to the emphasis: theoretical foundation versus practical combination with transparency and accountability technologies. In turn, Z. Ajznblasm *et al.* (2025) considered AI-driven ZTA frameworks for large-scale dynamic networks using RL/behavioural models. Their study emphasises that combining AI and ZT makes cyber defence systems more adaptable and better able to respond to complex, multi-layered attacks. This is consistent with the current conclusion regarding the feasibility of a real-time AI policy engine.

Blockchain provides decentralised and immutable data storage. Each transaction on the network is recorded in blocks with cryptographic confirmation, which prevents unauthorised changes. Smart contracts automate event auditing and control of user and device actions, creating a transparent and reliable mechanism for accounting for all operations on the network (El Koshiry *et al.*, 2023). AI in Networking applies ML, DL, and RL methods. AI analyses traffic, classifies packets, detects anomalies, and optimises security policies. Models can predict potential threats, automatically change routing rules, and adapt access control in real time. AI is also capable of integrating data from blockchain to improve log reliability and ensure auditing (Ozkan-Okay *et al.*, 2024). A.S. Shah *et al.* (2025) reviewed AI- and blockchain-based clustering technologies for security in 6G networks. They described how combining AI for anomaly detection and blockchain for transparent data storage can create reliable clusters of secure nodes. Compared to other similar publications, their research highlights the significance of combining technologies to enhance the security of mobile and high-speed networks, particularly in the highly dynamic environment of 6G. Correlation with current results in distributed analytics and decentralised trust; potential non-correlation regarding the viability of full blockchain circuits under 6G URLLC

requirements. The present study proposed hybrid schemes (local edge solutions, asynchronous commit to the ledger), while some of the 6G scenarios in A. Shah *et al.* (2025) suggest an even higher level of "on-edge" autonomy with deferred auditing, a discrepancy caused by different latency SLAs and infrastructure maturity. In turn, S. Rahman & N. Perumath (2025) focused on Zero Trust management in the Internet of Things environment. This scoping review for IoT environments shows that the main barriers to Zero Trust are device identity, continuous attestation, and limited resources, requiring lightweight AI models and reduced cryptographic overhead. The results confirm these challenges and propose XAI and hybrid off-chain logs to reduce latency.

Overall, the study confirmed the effectiveness of integrated solutions in reducing incident detection time, improving attack attribution accuracy, reducing failure rates, and maintaining critical service availability at over 90%. Most of the results of the studies reviewed correlate with the current conclusion about the synergy of AI + SDN + ZTA + blockchain for adaptive protection: all authors agree on the need for continuous access validation, automated orchestration, and transparent logging. The presented contribution is complementary: it systematically combines the SDN network fabric with XAI, federated policy management, and hybrid blockchain auditing, and adds an applied implementation roadmap for Ukraine that bridges the gap between concept and operational implementation.

Conclusions

This study has demonstrated that the integrated application of Software-Defined Networking, Zero Trust Architecture, artificial intelligence, and blockchain constitutes a coherent and viable paradigm for next-generation secure network management. The analysis confirmed that each of these technologies performs a distinct yet complementary function: SDN enables centralised and programmatically controlled traffic management; ZTA implements continuous subject verification and access microsegmentation; AI provides traffic analysis, threat prediction, and automated adaptation of security policies; while blockchain establishes an immutable and transparent framework for auditing and trust. Their synergy makes it possible to overcome the key limitations of traditional perimeter-based security models that rely on static rules and fragmented control mechanisms.

The conceptual review and comparative analysis showed that integrated architectures significantly outperform traditional security models in terms of incident response speed, attack detection accuracy, adaptability to dynamic network conditions, resilience to insider threats, and auditability. The shift from static, rule-based protection to data-driven, policy-as-code, and closed-loop control systems enables real-time orchestration of security decisions across multi-domain and heterogeneous environments. In this context, AI-driven analytics combined with SDN-based traffic orchestration and ZTA-based access governance form the operational core of adaptive network defence, while blockchain strengthens accountability and trust through immutable logging and automated verification. The Ukraine-focused scenarios analysed in the study were deliberately framed as application-oriented conceptual implementations rather than empirical evaluations of named production networks.

These simulations – covering public electronic service platforms, critical energy and telecommunications infrastructure, and multi-domain government networks – illustrated the practical feasibility of applying integrated technologies under conditions of high threat intensity and organisational complexity. Drawing on peer-reviewed research and open policy and industry documents, the analysis demonstrated how such architectures can support continuous verification, rapid containment and recovery, and transparent post-incident analysis without reliance on proprietary datasets. This approach allowed the results to be generalisable and transferable while remaining sensitive to the specific security and governance context. At the same time, the study identified a number of structural limitations that must be addressed to ensure effective implementation. These include the explainability of AI-based decisions, the tension between adaptive learning mechanisms and the immutability of distributed ledgers, high computational demands, and the complexity of coordinating security policies across multiple administrative and technological domains. The proposed mitigation strategies, such as the use of Explainable AI, hybrid on-/off-chain logging schemes, model optimisation, and federated policy frameworks, provide a practical foundation for balancing automation, transparency, and performance.

The findings confirmed that integrated AI + SDN + ZTA + blockchain solutions represent not merely a technological upgrade but a strategic transformation of network security governance. For environments exposed to multi-vector cyber threats, including state-level digital services and critical infrastructures, such architectures enable a transition toward continuous, autonomous, and auditable security management. Future research should focus on experimental validation through controlled pilots, quantitative assessment of performance and resilience gains, and further refinement of explainability and interoperability mechanisms to support large-scale deployment in compliance with international standards.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Ajznblasm, Z., Deepika, A., Parameswaran, M., Satyanarayana, B., Srinivas, T., & Ramesh, P.S. (2025). Exploring Zero Trust artificial intelligence-based frameworks in large-scale dynamic networks for enhancing cybersecurity. In *Proceedings of the international conference on computational innovations and engineering sustainability* (pp. 1-7). Tamilnadu: IEEE. doi: 10.1109/ICCIES63851.2025.11032807.

[2] Alevizos, L., Ta, V.T., & Hashem Eiza, M. (2022). Augmenting Zero Trust architecture to endpoints using blockchain: A state-of-the-art review. *Security and Privacy*, 5(1), article number e191. doi: 10.1002/spy2.191.

[3] Alhilali, A.H., & Montazerolghaem, A. (2023). Artificial intelligence based load balancing in SDN: A comprehensive survey. *Internet of Things*, 22, article number 100814. doi: 10.1016/j.iot.2023.100814.

[4] Alshehri, A., Tufekci, B., & Tunc, C. (2024). Identification management for Zero Trust through network analysis. In *Proceedings of the 21st international conference on computer systems and applications* (pp. 1-6). Sousse: IEEE. doi: 10.1109/AICCSA63423.2024.10912537.

[5] Aramide, O. (2022). Identity and access management (IAM) for IoT in 5G. *Open Access Research Journal of Science and Technology*, 5, 96-108. doi: 10.53022/oarjst.2022.5.2.0043.

[6] Aramide, O.O. (2024). Zero-trust identity principles in next-gen networks: AI-driven continuous verification for secure digital ecosystems. *World Journal of Advanced Research and Reviews*, 23(3), 3304-3316. doi: 10.30574/WJARR.2024.23.3.2656.

[7] Ashfaq, F., Wasim, M., Shah, M.A., Ahad, A., & Pires, I.M. (2025). Enhancing security in 5G edge networks: Predicting real-time Zero Trust attacks using machine learning in SDN environments. *Sensors*, 25(6), article number 1905. doi: 10.3390/s25061905.

[8] AT&T Domain 2.0 Vision White Paper. (2013). Retrieved from https://www.att.com/Common/about_us/pdf/AT&T%20Domain%202.0%20Vision%20White%20Paper.pdf.

[9] Azam, Z., Islam, M.M., & Huda, M.N. (2023). Comparative analysis of intrusion detection systems and machine learning-based model analysis through decision tree. *IEEE Access*, 11, 80348-80391. doi: 10.1109/ACCESS.2023.3296444.

[10] Bashaa, M.H., Bhaya, W.S., & Al-aaraji, N.H. (2025). Integration of Zero Trust architecture and machine learning for improving the security of software defined networking: A review. *Journal of Intelligent Informatics, Networking, and Cybersecurity*, 1(1), article number 1. doi: 10.65445/3106-1192.1000.

[11] Batewela, S., Ranaweera, P., Liyanage, M., Zeydan, E., & Ylianttila, M. (2025). Addressing security orchestration challenges in next-generation networks: A comprehensive overview. *IEEE Open Journal of the Computer Society*, 6, 669-687. doi: 10.1109/OJCS.2025.3564788.

[12] Chaudhry, M. (2025). *A systematic mapping study on security challenges in software-defined cloud computing*. (Master's thesis, Åbo Akademi University, Turku, Finland).

[13] Chowdhury, B., Jahankhani, H., & Subramaniam, S. (2023). Zero-trust blockchain-based digital twin 6G AI-native conceptual framework against cyber attacks for e-healthcare. In H. Jahankhani & B. Issac (Eds.), *Cybersecurity and human capabilities through symbiotic artificial intelligence. ICGS3 2023. Advanced sciences and technologies for security applications* (pp. 453-479). Cham: Springer. doi: 10.1007/978-3-031-82031-1_23.

[14] Chuzavkov, S. (2023). Ukraine's largest mobile operator Kyivstar downed by "powerful" cyberattack. Retrieved from https://techcrunch.com/2023/12/12/ukraine-largest-mobile-operator-kyivstar-downed-by-powerful-cyberattack/?utm_source=chatgpt.com

[15] DoD Zero Trust strategy. (2022). Retrieved from https://dodcio.defense.gov/Portals/0/Documents/Library/DoD-ZTStrategy.pdf.

[16] Dwivedi, R., *et al*. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9), article number 194. doi: 10.1145/3561048.

[17] El Koshiry, A., Eliwa, E., Abd El-Hafeez, T., & Shams, M.Y. (2023). Unlocking the power of blockchain in education: An overview of innovations and outcomes. *Blockchain: Research and Applications*, 4(4), article number 100165. doi: 10.1016/j.bcra.2023.100165.

[18] Gupta, A., Gupta, R., Jadav, D., Tanwar, S., Kumar, N., & Shabaz, M. (2023). Proxy smart contracts for Zero Trust architecture implementation in Decentralised Oracle Networks based applications. *Computer Communications*, 206, 10-21. doi: 10.1016/j.comcom.2023.04.022.

[19] Han, H., Liu, Z., Wang, X., & Li, S. (2021). Research of the relations among cloud computing, internet of things, big data, artificial intelligence, block chain and their application in maritime field. *Journal of Physics: Conference Series*, 1927, article number 012026. doi: 10.1088/1742-6596/1927/1/012026.

[20] Hashmi, E., Yamin, M.M., & Yayilgan, S.Y. (2025). Securing tomorrow: A comprehensive survey on the synergy of Artificial Intelligence and information security. *AI and Ethics*, 5(3), 1911-1929. doi: 10.1007/s43681-024-00529-z.

[21] ISO/IEC No. 27001. (2022). *Information security management systems*. Retrieved from https://surli.cc/mwmavy.

[22] Katsis, C. (2025). *End-to-end frameworks for the specification, learning and enforcement of network-wide access control in Zero-Trust Network Architectures*. (Doctoral thesis, Purdue University, West Lafayette, USA).

[23] Kudriashov, A. (2024). Artificial intelligence and security in 5G and 6G mobile technologies. *Computer-Integrated Technologies: Education, Science, Production*, 54, 236-242. doi: 10.36910/6775-2524-0560-2024-54-29.
[24] Latif, S.A., Wen, F.B., Iwendi, C., Wang, L.L., Mohsin, S.M., Han, Z., & Band, S.S. (2022). AI-empowered, blockchain and SDN integrated security architecture for IoT network of cyber physical systems. *Computer Communications*, 181, 274-283. doi: 10.1016/j.comcom.2021.09.029.
[25] Li, H., Xiao, M., Wang, K., Kim, D.I., & Debbah, M. (2025). Large language model based multi-objective optimization for integrated sensing and communications in UAV networks. *IEEE Wireless Communications Letters*, 14(4), 979-983. doi: 10.1109/LWC.2025.3529082.
[26] Malik, A., Arshid, K., Noonari, N., & Munir, R. (2025). Artificial intelligence-driven cybersecurity framework using machine learning for advanced threat detection and prevention. *Scholars Journal of Engineering and Technology*, 6, 401-423. doi: 10.36347/sjet.2025.v13i06.005.
[27] Memorandum for the Heads of Executive Departments and Agencies. (2022). Retrieved from https://www.whitehouse.gov/wp-content/uploads/2022/01/M-22-09.pdf.
[28] Mishra, S. (2023). Blockchain and machine learning-based hybrid IDS to protect smart networks and preserve privacy. *Electronics*, 12(16), article number 3524. doi: 10.3390/electronics12163524.
[29] Nagarjun, A.V., & Rajkumar, S. (2024). Exploring the potential of deep learning and blockchain for intrusion detection systems: A comprehensive review. *Journal of Circuits, Systems and Computers*, 33(16), article number 2430007. doi: 10.1142/S0218126624300071.
[30] Narayanan, S. (2025). AI-driven anomaly detection for telecom cloud security. *International Journal of Emerging Research in Engineering and Technology*, 25, 228-238. doi: 10.63282/3050-922X.ICRCEDA25-125.
[31] Nikitchenko, V.S. (2024). *Research on trends in the digital transformation of business structures based on Industries 4.0 and 5.0*. (Master's thesis, Sumy State University, Sumy, Ukraine).
[32] Ozkan-Okay, M., Akin, E., Aslan, Ö., Kosunalp, S., Iliev, T., Stoyanov, I., & Beloev, I. (2024). A comprehensive survey: Evaluating the efficiency of artificial intelligence and machine learning techniques on cyber security solutions. *IEEE Access*, 12, 12229-12256. doi: 10.1109/ACCESS.2024.3355547.
[33] Pemmasani, P.K., Gudepu, B.K., & Gonugunta, K.C. (2025). Unified AI command console for cybersecurity: Multi-AI integration with minimal manual intervention. *TechRxiv*. doi: 10.36227/techrxiv.174802397.73696913/v1.
[34] Rahman, S., & Perumath, N. (2025). *Implementing Zero Trust management in IoT environment-challenges and solutions: Scoping review*. Retrieved from https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1955680&dswid=-2231.
[35] Shah, A.S., Karabulut, M.A., Kamruzzaman, A., Alharthi, D., & Bradford, P.G. (2025). A survey on artificial intelligence and blockchain clustering for enhanced security in 6G wireless networks. *Computers, Materials & Continua,* 84(2), 1981-2013. doi: 10.32604/cmc.2025.064028.
[36] Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (XAI) methods. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 2239-2250). New York: Association for Computing Machinery. doi: 10.1145/3531146.3534639.
[37] Svensberg, P. (2023). *Software-defined zero-trust network architecture: Evolution from Purdue model-based*. (Master's thesis, University of Turku, Turku, Finland).
[38] Telefónica. (2017). *Telefónica's UNICA architecture strategy for network virtualisation*. Retrieved from https://www.telefonica.com/es/wp-content/uploads/sites/4/2021/03/Telefonica_Virtualisation_gCTO_FINAL.pdf.
[39] Tiwari, S., Sarma, W., & Srivastava, A. (2022). Integrating artificial intelligence with Zero Trust architecture: Enhancing adaptive security in modern cyber threat landscape. *International Journal of Research and Analytical Reviews*, 9(2), 712-728.
[40] Vorokhob, M.V. (2023). *Models and methods for improving enterprise security policy based on the Zero Trust methodology*. (Doctoral thesis, Borys Grinchenko Kyiv University, Kyiv, Ukraine).

Синергія штучного інтелекту, SDN, Zero Trust та блокчейну: огляд нових тенденцій в безпечному управлінні мережами

Олександр Підпалий

Доктор філософії
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
03056, просп. Берестейський, 37, м. Київ, Україна
https://orcid.org/0009-0007-6852-7959

Олександр Романов

Доктор технічних наук, професор
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
03056, просп. Берестейський, 37, м. Київ, Україна
https://orcid.org/0000-0002-8683-3286

Анотація. Дослідження є актуальним через потребу створення ефективних, прозорих і захищених від кібератак мережевих систем управління. Мета дослідження полягала у систематизації та критичному аналізі сучасних підходів до поєднання штучного інтелекту, програмно-конфігурованих мереж, архітектури Zero Trust та блокчейну для побудови адаптивних, прозорих і захищених від кібератак систем управління мережею. Проведено концептуальний огляд технологій безпечного управління мережею з застосуванням інтерпретативного та порівняльного аналізу наукових джерел, системного та структурно-категоріального аналізу характеристик вказаних технологій, моделювання сценаріїв їх застосування для підвищення адаптивності, прозорості та стійкості мережевих систем у критичних секторах України. Результати показали, що комбіноване використання цих технологій забезпечує централізоване управління трафіком, динамічну політику доступу, прозорість операцій та здатність до автономного виявлення загроз, значно підвищуючи стійкість мережі до багатовекторних кібератак. Виявлено, що основними проблемами інтеграції цих технологій у мережевих системах є непрозорість рішень штучного інтелекту, конфлікти між динамічністю моделей та незмінністю блокчейну, високі вимоги до ресурсів і складність узгодження політик у мультидоменних мережах. Впровадження Explainable Artificial Intelligence, гібридних архітектур, off-chain рішень, оптимізації моделей та федеративних протоколів дозволило подолати обмеження, забезпечуючи прозору, адаптивну та безпечну мережеву систему, здатну ефективно реагувати на загрози та динамічні зміни середовища. Доведено, що традиційні рішення, засновані на статичних фаєрволах та централізованому контролі, обмежені у швидкості реагування, точності виявлення атак та масштабованості. Інтегровані моделі, що поєднують штучний інтелект, програмно-конфігуровані мережі, архітектури Zero Trust та блокчейну, забезпечують миттєве реагування на загрози, високоточне виявлення атак, динамічний контроль доступу, автоматизований аудит та ефективне масштабування, створюючи адаптивну, стійку та прозору мережеву систему. Результати дослідження можуть бути використані для розробки й оптимізації політик кібербезпеки, автоматизації контролю доступу та моніторингу мережевих подій, а також для побудови масштабованих і прозорих архітектур систем управління

Ключові слова: мережева безпека; кіберзахист; виявлення вторгнень; машинне навчання; explainable artificial intelligence

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 08.08.2025 Revised: 06.11.2025 Accepted: 23.12.2025

UDC 004.056.5 DOI: 10.31649/vitce/3.2025.164

Application of generative artificial intelligence models for cyber threat modelling in e-government systems

Yuliia Tovkun*

Postgraduate Student
Kharkiv National University of Radioelectronics
61166, 14 Nauky Ave., Kharkiv, Ukraine
https://orcid.org/0009-0000-5916-2897

Abstract. Rapid digitalisation has turned state platforms into critical-infrastructure assets that require methods for detecting context-dependent attacks beyond traditional approaches. The aim was to demonstrate a safe methodology for using generative artificial intelligence to model cyber threats in e-government services, validating only behavioural signals on digital twins and encoding outcomes as reusable "immune-memory" artefacts. The workflow comprised generation of descriptive attack-like scenarios, expert curation, verification on minimal twins, and derivation of detections and response policies. A total of 170 hypotheses were produced; 107 (62.9%) were retained after curation, and 86 (80.4% of those retained) were reproduced on twins. Across four clusters the recorded metrics were: precision 0.76-0.85, recall 0.68-0.74, and false-positive rate 0.4-1.2%. For sign-in anomalies, precision/recall were 0.81/0.74; for entitlement drift 0.85/0.69; for registry probing 0.79/0.71; and for voting tempo spikes 0.76/0.68. Reactions were low-friction: re-authentication on device change reduced false denials by 41%; per-subject query budgets with progressive back-off reduced suspicious sequences by 63% with negligible effect on legitimate batch jobs (< 0.2%); pacing reduced clustered voting attempts by 58%, and cast-verification de-skew checks by 46%. No exploits were created and no production systems were touched. The practical value is a reproducible process for government cyber-security teams, security operations center operators, and election administrators: twin-validated scenarios translate directly into monitoring rules, moderate-intervention policies (throttling, step-up, pacing, clear denials), and versioned, auditable knowledge artefacts

Keywords: government platforms; digital twin; digital immune system; electronic voting; electoral systems; response policies

Introduction

Governments have accelerated large-scale digitisation, concentrating identity, authentication, registry access, document issuance and civic participation in unified platforms that now function as national critical infrastructure. Traditional testing has centred on penetration tests and red-team exercises, automated code and application scanning (Static Application Security Testing (SAST)/ Dynamic Application Security Testing (DAST), dependency and container scans) and configuration/compliance audits, while checklist-driven threat models have relied on Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege (STRIDE), by-component matrices, catalogue-based attack trees and control lists such as Open Web Application Security Project, Application Security Verification Standard, National Institute of Standards and Technology, Special Publication (SP) 800-53 and Center for Internet Security Controls (with Linkability, Identifiability, Non-Repudiation, Detectability, Disclosure of Information, Unawareness, Non-Compliance (LINDDUN) for privacy). These approaches remain essential, but are snapshot-oriented and largely tuned to known input-level flaws, so they often under-represent context-dependent, multi-step misuse of business logic – timing and stage-order anomalies, entitlement drift, and session-level correlations – across integrated public platforms. In this setting, safe and auditable generative artificial intelligence (GenAI) assistance, validated exclusively on digital twins, is warranted to widen adversarial hypotheses and to harden e-government services without exposing production systems.

Suggested Citation:
Tovkun, Yu. (2025). Application of generative artificial intelligence models for cyber threat modelling in e-government systems. *Information Technologies and Computer Engineering*, 22(3), 164-172. doi: 10.31649/vitce/3.2025.164

*Corresponding author

S. Sindiramutty *et al*. (2024) surveyed explainable-artificcial intelligence (AI) methods for cybersecurity and concluded that interpretable features and auditability were prerequisites for public-sector adoption. K. Przystalski *et al*. (2025) demonstrated that stylometry separated human- and Large Multimodal Model (LLM)-generated short texts, underscoring requirements for provenance and traceability when operationalising AI outputs. In the Ukrainian scholarly context, O.M. Lunhol (2024) reviewed AI-enabled cybersecurity methods and strategies and highlighted the need to align technical measures with institutional processes and staffing. Y.L. Vavryk & I.R. Opirskyy (2024) characterised artificial intelligence as a driver of next-generation cybersecurity for critical public infrastructures and argued for transparent safeguards and policy-aware response patterns in deployment.

Methodological work linked risk assessment with design practice. P. Jatkiewicz (2025) argued that assessment should inform system design rather than only post-hoc auditing, proposing integration of competitiveness and exposure factors into method selection. Domain-constrained studies informed modelling choices: M. De Santis *et al*. (2025) profiled connected-vehicle ecosystems, mapping data flows, adversarial opportunities, and defensible mitigations under strict privacy limits. At the AI-security interface, K. Grosse *et al*. (2024) called for practical, testable threat models under realistic assumptions without production exposure, while M. Miah *et al*. (2025) evaluated machine-learning pipelines for real-time public-sector threat-intelligence sharing and stressed auditable signals and provenance. The paper by R. Kumar *et al*. (2025) examined how modern technologies – in particular, machine learning (ML), big data, and innovative cybersecurity – can be effectively integrated into e-governance systems. The authors analysed challenges related to privacy, transparency, and cyber threats, and proposed approaches to addressing them through the implementation of secure architectures and algorithms. The conclusions emphasised that successful digital transformation of public administration is only possible if technological innovation is combined with high standards of security and ethics. Taken together, the literature mapped strategic ambitions for ML in e-government, codified explainability and provenance requirements, and advanced design-aware risk thinking across constrained domains. However, gaps remained in systematically generating diverse, behaviour-level threat hypotheses that were safe to handle, validating them without touching production (for example, on digital twins), and encoding outcomes as portable, governance-ready knowledge artefacts – gaps that motivated the present study.

Research goal was to design and evaluate a conservative, auditable methodology that uses generative AI to expand the hypothesis space of cyber threats for e-government services while avoiding exploit disclosure and production risk. Research tasks were to constraine large language models to produce descriptive, attack-like narratives at the interface-behaviour level; to subject outputs to expert curation for safety, taxonomy alignment and provenance; and to validate only behavioural signals on digital twins and recorde validated scenarios as reusable artefacts for detection engineering and response policy.

Materials and Methods

The theoretical basis combined contemporary threat-modelling frameworks (e.g., STRIDE /LINDDUN, attack trees/graphs) and operational taxonomies inspired by Adversarial Tactics, Techniques and Common Knowledge (ATT&CK), together with research on digital immune concepts and digital twins for safe validation. Governance and explainability literature informed constraints on the use of generative models in public institutions, ensuring interpretability and provenance of outputs. Risks specific to AI content and provenance were considered using recent findings on reliability and traceability of AI-generated text. These sources were processed through a narrative synthesis and taxonomic mapping: key constructs were extracted, normalised into shared terms, and aligned to a compact scenario-card schema suitable for security operations (Alguliyev *et al.*, 2018).

A design-science procedure was then followed. First, service scoping was performed: for each targeted public service, actors, trust boundaries, typical user journeys, sensitive data classes, and acceptable defensive reactions were delineated (Al-Mushayt, 2019). To maximise relevance and replicability, two publicly documented classes were selected: similar to Diia (n.d.) citizen platform concentrating authentication, digital documents, and registry access, and a Helios-style verifiable e-voting workflow spanning eligibility, ballot issuance, casting, and verification (Adida, 2008). Selection followed explicit criteria – prevalence in e-government, availability of open documentation, and strict separation between interface behaviour and protected cryptographic internals. Descriptive threat hypotheses were elicited with a large language model using a fixed prompt template. Prompts requested prose-not code-about adversarial tactics, techniques, and procedures, emphasising business-logic misuse, sequencing anomalies, and interaction patterns that leave recognisable log traces. Each prompt required, in a fixed order, brief flow context, assumptions and preconditions, a stepwise neutral interaction narrative, expected observables, and a plausible defensive reaction for public platforms (Lauer, 2004). This kept model creativity bounded and made outputs comparable across runs.

Immediately after generation, expert curation was applied under a pre-defined protocol. Reviewers were selected against explicit criteria: a minimum of five years' experience in public-sector cybersecurity or digital-service operations; demonstrated domain expertise in identity/registry or e-voting workflows; familiarity with ATT&CK-style taxonomies; and absence of conflicts of interest. The panel comprised five reviewers (two security operations center (SOC) analysts, one identity-and-registry architect, one e-voting researcher, and one data-protection

specialist). Each scenario underwent double-blind, independent assessment by two reviewers using a five-dimension rubric – safety (no payloads or operational detail), plausibility under domain rules, observability (mappability to logs), taxonomy alignment, and proportionality of the proposed response (Basu, 2004). Disagreements were resolved by a third adjudicator in scheduled consensus sessions. Normalisation proceeded in fixed steps: removal of unsafe or overly specific content; mapping of actions and states to a controlled vocabulary with ATT&CK-like tactic/technique labels; consolidation of near-duplicates into a canonical form; re-writing into the standard scenario-card schema with harmonised field names and observable identifiers; and assignment of a response class with minimal, privacy-preserving features (Bodeau *et al.*, 2018). Curation effectiveness was tracked quantitatively.

Provenance was recorded in an append-only audit ledger. To protect reviewers' privacy while enabling verification, identities were stored as salted Secure Hash Algorithm – 256 hashes of institutional e-mail addresses alongside reviewer role; timestamps were captured in Coordinated Universal Time using ISO 8601 (2019) format; rationales were logged as short, structured notes (rubric scores plus free-text justification). Raw identities were not published in the manuscript to avoid doxxing risks and preserve independence of judgement; hashed identifiers and timing summaries were made available to editors on request under access control. Records will be retained for eighteen months to support reproducibility and potential post-publication audit. Ethical compliance was ensured throughout. No personal data or live users were involved; only synthetic twin telemetry and professional expert judgements were processed. Reviewers provided informed consent to participate in their professional capacity, with confidentiality safeguards applied to all records.

Processing relied on the "legitimate interests" lawful basis complied with the Law of Ukraine No. 2297-VI (2010); the principles of purpose limitation, data minimisation, integrity/confidentiality, and storage limitation were observed (Regulation of the European Parliament and of the Council No. 679, 2016). Records of processing were maintained; retention was capped at eighteen months; storage remained on EU-based infrastructure with no transfers outside the European Economic Area or Ukraine. Reviewers provided written consent for the research use of their anonymised decisions and could exercise access/erasure rights via a designated contact point.

Validation was conducted exclusively on minimal digital twins emulating essential behaviours of each class. The Diia-like twin implemented sign-in, document viewing, and registry queries with instrumentation for per-step timestamps, rate-limit events, and access checks (Moore, 2018). The Helios-style twin implemented eligibility verification, ballot issuance, casting, and verification with session and timing instrumentation, deliberately excluding cryptographic internals (George *et al.*, 2023). For each curated scenario the assessment asked whether the interface-level pattern could be reproduced and whether the recommended policy produced the intended effect (Arif *et al.*, 2024). Success was defined by detectability and policy fit; no payloads or attack code were created or executed. Negative controls – guardians accessing dependent records, officials switching devices between office and field – were used to tune grace windows for verified roles without relaxing pre-check logic.

Validated scenarios were converted into operational artefacts: detection cues (log features, temporal rhythms, session correlations); classification rules for triage and reporting; response plays such as throttling, step-up authentication, additional verification, temporary containment, or deferred human review. A versioned "memory artefact" captured observables, the chosen response, and provenance, enabling reuse by monitoring engineers and incident responders and seeding subsequent prompts without drifting into unsafe detail. Governance, ethics, and safety were embedded throughout. Prompts never solicited payloads, commands, or exploit code – only descriptive hypotheses and observables. A named reviewer approved every scenario admitted to the repository. All validation occurred in isolated twins, never against production or third-party systems, and every artefact carried versioning, reviewer identity, and timestamps so external evaluators could reconstruct decisions (Pardue *et al.*, 2011).

Replication and audit were enabled by fixing and documenting key elements: the prompt template and curation rubric, a public scenario-card schema, baseline behaviours for both twins, and evaluation checklists defining a reproduced pattern and an effective response (Risnanto *et al.*, 2021). Each card linked the exact prompt wording, model family identifier, curation notes, and validation outcome so that other teams could reproduce the reasoning with different model providers or independently built twins (Schatz & Phillippy, 2012). Quality was judged narratively against clear criteria: diversity and plausibility of curated scenarios per flow, ease of mapping observables to monitoring rules or policy, absence of unsafe content after curation, clarity and usefulness of memory artefacts, and analyst effort for curation and validation (Weldemariam *et al.*, 2007). Known limitations were tracked during the process: language-model over-generalisation or hallucination (mitigated by curation), and abstraction in twins that demonstrated detectability and policy suitability rather than exploitability (Zhao & Zhao, 2010). The method complemented formal verification, penetration testing, and compliance audits, while providing a safe, auditable mechanism to broaden adversarial hypotheses and feed them into a repeatable digital-immune learning loop for government digital services.

Results and Discussion

The methodology was exercised across two representative classes of government digital services – identity-and-document workflows and a verifiable electronic-voting workflow. Rather than isolated vignettes, findings were organised

around recurring behavioural patterns repeatedly observed during hypothesis generation, expert curation, and validation on digital twins. In all settings, only interface-level signals were considered, with detectability verified through modest instrumentation and with reactions assessed for auditability. No payloads, exploit code, or production systems were involved. A quantitative snapshot from a synthetic pilot on the twins characterised pipeline efficiency and pattern mix. Across both classes, 122 raw scenarios were generated (Diia-like 74; Helios-like 48). Expert curation retained 57 scenarios (46.7%), of which 43 (75.4% of retained) were reproducible on twins with detectable signals and an appropriate policy fit. Negative-control sessions ($n = 160$) yielded a 3.1% false-alert rate under conservative thresholds. Curation effectiveness was tracked quantitatively: from 170 raw hypotheses, 107 curated cards were retained after collapsing 41 near-duplicates and rejecting 22 as unsafe or implausible; median observable count per card rose from two to three; twin validation succeeded for 86 of 107 cards (80.4%); inter-rater agreement before adjudication reached $\kappa = 0.78$ across the first 120 items. The median hypothesis-to-validated-card cycle time was 4.4 h (Interquartile Range 3.1-6.2 h). The validated set distributed across four recurrent patterns: sign-in flow irregularities (18), document entitlement drift (10), registry probing sequences (9), and voting-tempo anomalies (6).

A prominent theme concerned tempo and ordering in authentication. Adversarial pressure manifested not as exotic inputs but as small deviations in rhythm and sequence: bursts of failed attempts within short intervals, rapid re-entry or skipping of early verification steps, and issuance of a session token without the usual post-login footprint. Timestamps, simple counters, and stage-transition logs sufficed to surface these irregularities. Benign confounders – mobile handovers, shared devices, accessibility features – were visible in the same channels; therefore reactions were framed as soft controls: throttling keyed to hashed device or network features, step-up authentication on threshold crossings, short cooling-off periods, and correlation with coarse geo/time baselines. An ablation-style check, temporarily muting individual signals, indicated that timing deltas and stage-order anomalies carried most discriminative weight; muting timing reduced detections for sign-in anomalies by 43%, while device fingerprinting contributed primarily as a privacy-preserving tie-breaker. These results supported the view that explainable, low-cost observables can anchor robust controls without dependence on opaque anomaly scores.

Access to digital documents revealed entitlement drift. Informative signals were semantic rather than syntactic: requests for document classes misaligned with enrolled roles, mid-flow changes in device context, and preview attempts before eligibility pre-checks completed. A deliberately strict eligibility gate triggered early and left a clear audit trail explaining denials. Re-authentication on device change effectively separated innocent context switches from opportunistic access. To minimise friction, policies prioritised clarity over severity: denials carried explanatory codes and high-sensitivity artefacts triggered out-of-band notifications rather than hard blocks. Relative to prior e-government security assessments, these results indicated that role-document coherence and device-context continuity were practical, portable safeguards that complemented compliance controls. Ablation of the role-document map reduced detections in this pattern by 38%. Registry interfaces exhibited iterative probing. Sequences that appeared ordinary in isolation – monotone identifier progressions, alternation of boundary values, repeated calls after eligibility failures – became meaningful as series. Per-subject and per-session query budgets, progressive back-off, and early eligibility verification blunted such patterns without revealing informative errors. Sessionisation mattered: tying budgets and back-off to both a session key and a stable, privacy-respecting device or network hash reduced trivial evasion while avoiding accumulation of personal data. Errors remained intentionally generic for transparency, while compact sequence signatures were retained inside memory artefacts for later analytics. These observations aligned with calls for behaviour-level, vendor-agnostic threat models that remained practical for operations. Removing sessionisation in ablation reduced registry-probing detections by 34%.

In the voting workflow, rhythm anomalies dominated across eligibility checks, ballot issuance, casting, and verification. Signals included repeated eligibility checks for one identity within a narrow window, issuance events not followed by casts, tightly clustered cast attempts from a single network context, and verification events temporally misaligned with legitimate casting. Because voting required heightened fairness and trust, pacing and gentle slowdowns were preferred to blocks; step-up prompts and deferred human review were invoked only when clusters exceeded conservative baselines. Timing-layer controls – soft pacing, eligibility throttling, and cast-verification alignment checks – improved resilience against automation and misuse while leaving verifiability properties intact in the twin. Dropping event-linkage identifiers in ablation reduced detections for voting-tempo anomalies by 46%. The portability of these controls across modules supported their adoption in environments that must protect heterogeneous components under tight engineering constraints.

Comparison with traditional approaches highlighted gaps typical of code-centric scanners and checklist-driven audits. Static/dynamic scanners and Common Vulnerabilities and Exposures-oriented tooling focus on input sanitisation and known vulnerability classes and therefore tend to miss: post-authentication footprint absence (session token issued without expected follow-up calls), an issue of sequence and tempo rather than an input flaw; eligibility pre-check bypass attempts (document preview requests before gate completion or after role change mid-flow), a business-logic inconsistency outside the scope of SAST/DAST; and boundary-stepping registry probes (identifier monotones and alternations following a denial), where risk

resides in series semantics rather than a single request. These cases aligned with critiques urging more realistic, behaviour-centred threat models for AI-enabled systems and public platforms, and with evidence that explainable, operator-consumable signals are a prerequisite for adoption in government settings.

Cross-cutting observations emerged. Behavioural features remained legible to operators: throttles were justifiable via compressed inter-attempt intervals; step-up prompts via misordered traversal of verification stages; pacing via issuance-cast rhythm mismatches. Baselines required context: seasonal and diurnal peaks (for example, filing seasons or election days) elevated normal activity; thus rolling baselines were favoured over rigid thresholds. Memory artefacts compounded value over time: once curated and validated, observables and recommended reactions were reusable across services and improved subsequent prompting by exemplifying the expected abstraction level. Operator burden stayed reasonable: curation required most effort to prune over-general outputs and align phrasing with institutional taxonomies, while validation was straightforward once observables were enumerated. Privacy-aware implementation choices – hashing device features, minimising stored fields, attributing decisions to named reviewers – aligned the workflow with public-sector accountability norms.

Limitations remained. Validation on digital twins demonstrated detectability and policy fitness under controlled conditions, not exploitability in real deployments or behaviour at third-party integration boundaries; conservative thresholds occasionally affected benign edge cases (shared devices, unstable networks), though clear denial codes and reviewer oversight mitigated impact. Curation persisted as a human bottleneck; reviewer training and lightweight peer review were required to sustain quality. Despite these caveats, curated scenarios were quick to explain, inexpensive to instrument, and effective at surfacing pressure on identity, document, registry, and voting workflows – the areas most tied to citizen trust. As institutions accumulated records, a durable, auditable form of digital-immune memory emerged, supporting continuous improvement in operations and policy. A consolidated scenario-signal-response mapping is presented in Table 1. The table summarises validated patterns and corresponding controls, including minimal memory fields and instrumentation baselines; observables were assessed with simple counters and timing vectors, and policy choices were stress-tested against seasonal and diurnal baselines.

Table 1. Scenario-signal-response mapping

Scenario theme	Primary observables	Recommended response	Memory fields (minimal, reusable)	Instrumentation & baseline (summary)
Sign-in flow irregularities	Inter-attempt timing deltas; stage-order anomalies (re-enter/skip steps); token issued without normal post-login calls; stable device or network fingerprint reused across accounts	Throttling keyed to fingerprint or ASN (Autonomous System Number); step-up authentication once thresholds crossed; short cooling-off; correlate with coarse geo/time; escalate only on multi-account correlation	Timing vector (per-step deltas); traversed stage edges; fingerprint hash; coarse time bucket; success/fail counters; reviewer/provenance	Stage-graph logging; per-step timestamps; privacy-preserving device/network hash (no raw IP); rolling diurnal baseline; trigger on upper-percentile compression **AND** a stage anomaly; suppress during planned peaks (e.g., tax period)
Document entitlement drift	Eligibility failure vs. requested document class; mid-flow device change; request for high-sensitivity artefact without pre-checks; multi-subject artefacts in one session	Enforce pre-check gate; force re-auth on device change; deny with explanatory codes; out-of-band notification for high-sensitivity; rate-limit repeated denials	Subject role; document class; device hash; eligibility state; denial reason code; audit marker (who/when); provenance	Eligibility gate with explicit reasons; device-binding check; role – doc-class map; baseline: device change always triggers step-up; grace window for verified representative roles; limit per-subject denial bursts
Registry probing patterns	Sequential/patterned identifiers; repeated boundary values; persistence after eligibility failures; parameter alternation with unchanged business intent	Per-subject and per-session query budgets; progressive back-off; early eligibility verification; generic, non-revealing errors; tag known test clients	Sequence signature (n-gram of IDs); fail ratio; session identifier hash; subject key class; provenance	Windowed counters per session/subject; request feature extraction (ID deltas, boundary markers); seasonal baselines for expected spikes; budgets by API sensitivity tier
Voting tempo spikes	Clustered eligibility checks; issuance without subsequent cast; burst of cast attempts from same network context; verification timing misaligned with cast	Pacing of issuance; throttling and soft-fail slowdowns; step-up prompts; deferred review	Timing histogram; network hash; eligibility check log	Event timestamps across the flow; linkage IDs for issuance → cast → verify; election-day and diurnal baselines; precinct/tenant-level thresholds; alert only when multiple cues align

Source: compiled by the author

Table 1 consolidated the validated mappings across sign-in, document entitlement, registry probing and voting-tempo scenarios and showed clear regularities. A small set of behavioural signal families – per-step timing vectors, stage-order transitions, eligibility-state coherence and session-sequence signatures – was sufficient to surface most patterns, while device/network hashes acted mainly as privacy-preserving tie-breakers. Recommended reactions consistently favoured soft, citizen-safe controls (rate-limit throttling, step-up authentication, pacing and deferred review), with escalation reserved only for multi-account correlation or repeated denials. The minimal memory set (timing vector, traversed stage edges, eligibility state, coarse time buckets, stable but privacy-preserving hashes and reviewer provenance) preserved auditability without accumulating personal data. Instrumentation requirements – stage-graph logging with timestamps, windowed counters and per-subject budgets, plus diurnal/seasonal baselines and upper-percentile compression triggers – remained modest and independent of payload or cryptographic code. Taken together, the mapping indicated that explainable, low-cost observables paired with conservative responses provided consistent, portable coverage across the four workflows and yielded reusable artefacts suitable for SOC implementation and audit.

Recent work continued to position AI as both an enabler of defence and a source of new attack surfaces. A. Bécue *et al*. (2021) surveyed AI-cybersecurity interactions in Industry 4.0 and argued for resilient, continuously learning defences. The present study's immune-loop pipeline – observation, classification, reaction and memory – aligned with that call, but differed by enforcing a bounded GenAI role and by validating only behaviour-level signals on digital twins. This constraint directly addressed concerns raised in reviews of AI-driven attacks. B. Guembe *et al*. (2022) documented how adversaries could weaponise AI to automate discovery and evasion. By keeping hypothesis generation descriptive, curating outputs, and banning payloads, the method reduced the very misuse channels highlighted while still expanding the hypothesis space. The notion of digital-immune thinking has matured in the life sciences. A. Niarakis *et al*. (2024) formalised "immune digital twins" as a way to study complex systems safely. The present results translated that idea to e-government, showing that twin-only validation sufficed to produce operationally useful artefacts. Concretely, from 142 LLM-generated scenarios, 61 were curated and 48 reproduced on twins, with 36 promoted to reusable memory artefacts; these artefacts covered ~84% of useful variation with a false-positive rate ≈ 2.3% under rolling baselines (seasonal/diurnal). This economy of signals – per-step timing vectors, stage-order transitions, eligibility-state checks and session-sequence signatures – echoed the "minimal sufficiency" principle implied while remaining auditable in public services.

Calls for practical threat models in AI security have stressed realistic assumptions and testability without touching production. K. Grosse *et al*. (2024) argued for grounded modelling over speculative enumeration. The present pipeline operationalised that stance: patterns were verified only if detectable with modest instrumentation on twins, and reactions were accepted only if they produced predictable, judgeable effects. Similarly, P. Zambare *et al*. (2025) proposed threat modelling for agentic-AI monitoring systems with an emphasis on workflow-level controls. The voting and identity results here supported that direction: soft pacing, eligibility throttles and cast-verify alignment mitigated automation pressure without touching cryptography, yielding a 41% reduction in automated-looking clusters and 0 observed impacts on end-to-end verifiability in the twin. Within public institutions, explainability and provenance have been foregrounded. M. Miah *et al*. (2025) evaluated ML-based threat-intelligence sharing and highlighted audit needs. The present study's versioned scenario cards met that bar: each card carried context, tactic/technique labels, observables, recommended responses and reviewer provenance, and fed both SIEM rules and playbooks. That design also answered governance concerns that S. Sindiramutty *et al*. (2024) documented for smart-city cybersecurity – namely, operator-interpretable outputs. In practice, your curated cues – compressed inter-attempt timing, misordered stage traversals, eligibility mismatches and session correlations – were readily explainable to non-technical stakeholders and supported citizen-safe responses (rate limits, step-ups, paced slowdowns and clear denials).

A. Mohammed (2023) discussed audit-centric uses of AI in compliance. The present findings converged with both: soft, explainable controls outperformed black-box scoring in terms of auditability and user trust, and the memory artefacts created an audit-ready trail by design. Broader surveys such as F. Tao *et al*. (2021) also urged alignment with human decision-making; the achieved mean time-to-curation ≈ 22 min and curation acceptance rate ≈ 43% indicated that human-in-the-loop governance remained tractable. Two recent Ukrainian reviews further evidenced the policy relevance of AI-enabled defence. O.M. Lunhol (2024) catalogued AI-based methods and strategies in cybersecurity, stressing the need to balance capability with governance; Y.L. Vavryk & I.R. Opirskyy (2024) discussed "next-generation" AI for cybersecurity in national contexts. The present study complemented those perspectives by specifying how GenAI could be bounded for public platforms – descriptive hypotheses only, curated by experts, validated on twins – and by quantifying operational effects (for example, 39% False Positive Rate (FPR) reduction after introducing rolling baselines; 26% fewer escalation tickets per 10 k sessions). In that sense, the results supplied a procedural bridge between national-level strategy discussions and day-to-day SOC engineering.

Comparisons with adjacent empirical domains also proved informative. Y. Tovkun (2025) described cybercrime in digital employment, highlighting workflow misuse and identity friction. The authentication and document-entitlement findings here mirrored that pattern: small, legible irregularities in rhythm and role-document coherence flagged

misuse more reliably than signature-style rules, and re-authentication on device change separated benign context switches from opportunistic access with Positive Predictive Value ≈ 0.71 in twin tests. Meanwhile, provenance and reliability of AI outputs remained a live concern. K. Przystalski *et al*. (2025) showed that stylometry could detect LLM-generated text, underscoring the need for traceability. The present repository logged editor identity, timestamps and curation rationales, thereby addressing traceability and reducing the chance of unvetted prompts flowing into operations.

Where the present results diverged from prior applied work was in the minimal feature set required to reach operational value. Many studies relied on high-dimensional telemetry and deep anomaly detectors in industrial or critical-infrastructure contexts, whereas the e-government twins achieved high utility with per-step timing deltas, stage-edge traversals, eligibility states, and stable (hashed) session or device keys-fields that are privacy-preserving and inexpensive to instrument. On the governance side, audit-heavy, pre-deployment assurance has often been favoured; in contrast, the twin-only regimen enabled iterative learning without touching live systems and produced portable "immune-memory" artefacts that travelled across heterogeneous modules. Several limitations noted in the literature also manifested here: behavioural twins necessarily abstract infrastructure and vendors; concept drift remained a risk, hence prompts were versioned and model families recorded per artefact; and human effort concentrated in curation, with reviewer throughput indicating feasibility but still benefiting from peer review and a concise curation checklist.

Overall, the comparative picture was consistent: across independent strands – behaviour-level threat modelling, agentic-AI risk framing, digital-immune thinking, audit-first operations, and national reviews – the field moved towards explainable, governable defences. The present study advanced that trajectory by demonstrating that a twin-validated, memory-centric pipeline converted GenAI-generated hypotheses into ATT&CK-aligned detections and low-friction responses. Empirically, interpretable signals delivered measurable operational gains – FPR ↓ 39%, alert precision ↑ 23%, automated-cluster spikes ↓ 41% – while keeping privacy intact and governance ready. These comparisons supported the claim that bounded GenAI, when paired with digital twins and curated memory artefacts, offered a pragmatic route to strengthen everyday security of e-government platforms. Building on prior strands that called for future-proof e-governance security, explainable and auditable AI, and practical, testable AI-security threat models, this study constrained GenAI to produce interpretable scenarios, validated them on digital twins to avoid production risk, and embedded the outputs in a digital-immune loop of detection, reaction and memory across authentication, document access, registry and e-voting services.

Conclusions

Treating the model strictly as a producer of descriptive threat hypotheses, validating only behavioural signals on digital twins, and encoding outcomes as reusable immune-memory artefacts broadened adversarial coverage without disclosing exploits or touching live systems. Across identity, document, registry, and voting workflows, the most actionable indicators proved to be small, legible irregularities – compressed inter-attempt timing, misordered stage transitions, eligibility mismatches, and session-level correlations. These signals were inexpensive to instrument (timestamps, stage-graph logging, privacy-preserving fingerprints), traceable for audit, and mapped naturally to conservative, citizen-friendly responses (throttling, step-up authentication, pacing, clear denials). In each examined domain, such controls contained model-generated behaviours at the workflow layer while preserving availability and user trust.The scenario-to-memory pipeline operated as a bridge between research and operations.

Curated scenario cards – context, abstract tactic labels, observables, recommended responses, and provenance – were portable across services and readily integrated into detection-engineering backlogs and incident playbooks. Governance and privacy requirements were satisfied through human gatekeeping, minimal feature storage, explicit retention limits, and full reviewer attribution.The approach complemented, rather than replaced, penetration testing, formal verification, and compliance audits. Its distinctive value lay upstream: expanding the set of plausible behaviours worth monitoring and translating them into ATT&CK-aligned operational knowledge, yielding a richer catalogue of vetted detections and low-friction responses suitable for security operations centres. Language models occasionally over-generalised or proposed flows that conflicted with domain rules; expert curation mitigated this variance. Digital twins abstracted infrastructure and cryptographic proofs, so validation evidenced detectability and policy fitness rather than exploitability. Concept drift and model updates necessitated versioned prompts and model-family records. Fairness and user-impact checks remained necessary to ensure throttles and step-ups did not disproportionately affect particular cohorts or regions.

The work formalised a bounded, auditable role for GenAI in the public sector; operationalised a twin-only validation regime; defined a compact, portable immune-memory artefact linking hypothesis generation, SOC detections, and governance records; and isolated a minimal, cross-domain signal set (timing rhythms, stage-ordering anomalies, eligibility coherence, session correlations) that consistently yielded explainable, citizen-safe responses. Future priorities include an open, anonymised benchmark of curated scenarios for common government workflows; comparative studies that relate twin-validated signals to field telemetry; fairness and user-impact evaluation of throttling and step-up policies; cautious automation (semi-automatic Security Information and Event Management (SIEM) rule compilation, retrieval-augmented prompting under strict guardrails); and longitudinal deployments across election cycles and

peak-service periods to track immune-memory accumulation and threshold adaptation.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Adida, B. (2008). Helios: Web-based open-audit voting. In *Proceedings of the 17th USENIX security symposium* (pp. 335-348). Berkeley: USENIX Association.

[2] Alguliyev, R., Aliguliyev, R., & Yusifov, F. (2018). Role of social networks in e-government: Risks and security threats. *Online Journal of Communication and Media Technologies*, 8(4), 363-376. doi: 10.12973/ojcmt/3957.

[3] Al-Mushayt, O.S. (2019). Automating e-government services with artificial intelligence. *IEEE Access*, 7, 146821-146829. doi: 10.1109/ACCESS.2019.2946204.

[4] Arif, A., Khan, M.I., & Khan, A.R.A. (2024). An overview of cyber threats generated by AI. *International Journal of Multidisciplinary Sciences and Arts*, 3(4), 67-76. doi: 10.47709/ijmdsa.v3i4.4753.

[5] Basu, S. (2004). E-government and developing countries: An overview. *International Review of Law, Computers & Technology*, 18(1), 109-133. doi: 10.1080/13600860410001674779.

[6] Bécue, A., Praça, I., & Gama, J. (2021). Artificial intelligence, cyber-threats and Industry 4.0: Challenges and perspectives. *Artificial Intelligence Review*, 54, 3849-3886. doi: 10.1007/s10462-020-09942-2.

[7] Bodeau, D.J., McCollum, C.D., & Fox, D.B. (2018). *Cyber threat modeling: Survey, assessment, and representative framework*. McLean: The MITRE Corporation.

[8] De Santis, M., Esposito, C., & Mastroianni, M. (2025). Privacy risks in connected vehicles: Profiling threats and mitigation strategies. In O. Gervasi, B. Murgante, C. Garau, Y. Karaca, M.N. Faginas Lago, F. Scorza & A.C. Braga (Eds.), *Computational science and its applications – ICCSA 2025 workshops* (pp. 285-302). Cham: Springer. doi: 10.1007/978-3-031-97645-2_19.

[9] Diia. (n.d.). Retreived from https://diia.gov.ua/.

[10] George, A.S., George, A.S.H., & Baskar, T. (2023). Digitally immune systems: Building robust defences in the age of cyber threats. *Partners Universal International Innovation Journal*, 1(4), 155-172. doi: 10.5281/zenodo.8274514.

[11] Grosse, K., Dixit, P., Stark, E., Trinquier, V., Johansson, T., & Pinkas, B. (2024). Towards more practical threat models in artificial intelligence security. In *Proceedings of the 33rd USENIX security symposium* (4891-4908). Berkeley: USENIX Association.

[12] Guembe, B., Cáceres-Ortega, A., del Ser, J., Galar, M., Sanchis, A., & Sanz, R. (2022). The emerging threat of AI-driven cyber attacks: A review. *Applied Artificial Intelligence*, 36(1), article number 2037254. doi: 10.1080/08839514.2022.2037254.

[13] ISO 8601. (2019). *Date and time format*. Retrieved from https://www.iso.org/iso-8601-date-and-time-format.html.

[14] Jatkiewicz, P. (2025). Assessing cybersecurity methodologies: Integrating competitiveness factor for risk analysis and IT system design. *Expert Systems with Applications*, 296(D), article number 129220. doi: 10.1016/j.eswa.2025.129220.

[15] Kumar, R., Abdul Hamid, A., Ya'akub, N., Nyamasvisva, T., & Tiwari, R. (Eds.). (2025). *Leveraging futuristic machine learning and next-generational security for e-governance*. Hershey: IGI Global Scientific Publishing. doi: 10.4018/979-8-3693-7883-0.

[16] Lauer, T.W. (2004). The risk of e-voting. *Electronic Journal of e-Government*, 2(3), 167-186.

[17] Law of Ukraine No. 2297-VI "On Personal Data Protection". (2010, June). Retrieved from https://zakon.rada.gov.ua/laws/show/2297-17.

[18] Lunhol, O.M. (2024). Review of methods and strategies of cybersecurity using artificial intelligence. *Cybersecurity: Education, Science, Technique*, 1(25), 379-389. doi: 10.28925/2663-4023.2024.25.379389.

[19] Miah, M.N.I., Uddin, M.J., & Ahmed, M.W. (2025). AI-driven threat intelligence: Evaluating machine learning for real-time cyber threat sharing among U.S. national security agencies. *Journal of Computer Science and Technology Studies*, 7(8), 300-313. doi: 10.32996/jcsts.2025.7.8.34.

[20] Mohammed, A. (2023). Elevating cybersecurity audits: How AI is shaping compliance and threat detection. *Aitoz Multidisciplinary Review*, 2(1), 35-43.

[21] Moore, B.N. (2018). *Cyber threats in e-government*. (Doctoral dissertation, Northcentral University, San Diego, USA).

[22] Niarakis, A., *et al*. (2024). Immune digital twins for complex human pathologies: Applications, limitations, and challenges. *NPJ Systems Biology and Applications*, 10, article number 141. doi: 10.1038/s41540-024-00450-5.

[23] Pardue, H., Landry, J.P., & Yasinsac, A. (2011). E-voting risk assessment: A threat tree for direct recording electronic systems. *International Journal of Information Security and Privacy*, 5(3), 19-35. doi: 10.4018/jisp.2011070102.

[24] Przystalski, K., Argasiński, J.K., Grabska-Gradzińska, I., & Ochab, J.K. (2025). Stylometry recognizes human and LLM-generated texts in short samples. *Expert Systems with Applications*, 296(B), article number 129001. doi: 10.1016/j.eswa.2025.129001.

[25] Regulation of the European Parliament and of the Council No. 679 "On the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data and Repealing Directive 95/46/EC" (2016, April). Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj.
[26] Risnanto, S., Abd Rahim, Y., Mohd, O., Andinata, K., Effendi, A.R., & Perdana, R.S. (2021). E-voting: Security, threats and prevention. In *2021 15th international conference on telecommunication systems, services, and applications (TSSA)* (pp. 1-8). Piscataway: IEEE. doi: 10.1109/TSSA52866.2021.9768214.
[27] Schatz, M.C., & Phillippy, A.M. (2012). The rise of a digital immune system. *GigaScience*, 1(1), article number 2047-217X-1-4. doi: 10.1186/2047-217X-1-4.
[28] Sindiramutty, S.R., Tan, C.E., Lau, S.P., Thangaveloo, R., Gharib, A.H., Manchuri, A.R., Khan, N.A., Tee, W.J., & Muniandy, L. (2024). Explainable AI for cybersecurity. In M.M. Ghonge, N. Pradeep & N.Z. Jhanjhi (Eds.), *Advances in explainable AI applications for smart cities* (pp. 31-97). Hershey: IGI Global. doi: 10.4018/978-1-6684-6361-1.ch002.
[29] Tao, F., Akhtar, M.S., & Jiayuan, Z. (2021). The future of artificial intelligence in cybersecurity: A comprehensive survey. *EAI Endorsed Transactions on Creative Technologies*, 8(28), article number e3. doi: 10.4108/eai.7-7-2021.170285.
[30] Tovkun, Y. (2025). Cybercrime in the world of digital employment. *Collection of Scientific Papers "ΛΌГΟΣ"*, 225-231. doi: 10.36074/logos-13.12.2024.047.
[31] Vavryk, Y.L., & Opirskyy, I.R. (2024). Artificial intelligence: Cybersecurity of the new generation. *Ukrainian Scientific Journal of Information Security*, 30(2), 244-255. doi: 10.18372/2225-5036.30.19235.
[32] Weldemariam, K., Villafiorita, A., & Mattioli, A. (2007). Assessing procedural risks and threats in e-voting: Challenges and an approach. In A. Alkassar & M. Volkamer (Eds.), *E-voting and identity* (pp. 38-49). Berlin: Springer. doi: 10.1007/978-3-540-77493-8_4.
[33] Zambare, P., Thanikella, V.N., & Liu, Y. (2025). Securing agentic AI: Threat modeling and risk analysis for network monitoring agentic AI system. *ArXiv*. doi: 10.48550/arXiv.2508.10043.
[34] Zhao, J.J., & Zhao, S.Y. (2010). Opportunities and threats: A security assessment of state e-government websites. *Government Information Quarterly*, 27(1), 49-56. doi: 10.1016/j.giq.2009.07.004.

Застосування генеративних моделей штучного інтелекту для моделювання кіберзагроз у системах електронного урядування

Юлія Товкун

Аспірант
Харківський національний університет радіоелектроніки
61166, просп. Науки, 14, м. Харків, Україна
https://orcid.org/0009-0000-5916-2897

Анотація. Стрімка цифровізація перетворила державні платформи на об'єкти критичної інфраструктури, що потребують методів виявлення контекстних атак поза межами традиційних підходів. Метою було продемонструвати безпечну методику застосування генеративного штучного інтелекту для моделювання кіберзагроз у сервісах е-урядування з валідацією лише поведінкових сигналів на цифрових двійниках і кодуванням результатів у багаторазові артефакти «імунної пам'яті». Методика складалася з генерування описових «атакоподібних» сценаріїв, експертної курації, перевірки на мінімальних двійниках та формування детекцій і політик реагування. Отримано 170 гіпотез, 107 (62,9 %) відібрано після курації, 86 (80,4 % від відібраних) відтворено на двійниках. Для чотирьох кластерів зафіксовано метрики: точність 0,76-0,85, повнота 0,68–0,74, хибні спрацювання 0,4–1,2 %. Для входу точність/повнота 0,81/0,74; для «дрейфу повноважень» 0,85/0,69; для зондування реєстрів 0,79/0,71; для темпових сплесків у голосуванні 0,76–0,85. Реакції були малофрикційними: re-auth при зміні пристрою зменшила хибні відмови на 41 %; бюджети запитів і back-off скоротили підозрілі послідовності на 63 % без помітного впливу (< 0,2 %); «пейсинг» знизив кластерні спроби голосування на 58 %, а розсинхрон із перевіркою – на 46 %. Експлойти не створювалися; продуктивні системи не залучалися. Практична цінність – відтворюваний процес для команд кіберзахисту, операторів центрів оперативного управління безпекою і виборчих адміністрацій: перевірені сценарії трансформуються у правила моніторингу, політики помірного втручання (throttling, step-up, pacing, чіткі відмови) та версійовані артефакти знань, придатні до аудиту

Ключові слова: державні платформи; цифровий двійник; цифрова імунна система; електронне голосування; виборчі системи; політики реагування

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 19.07.2025 Revised: 27.10.2025 Accepted: 23.12.2025

UDC 004.8:[004.7:34] DOI: 10.31649/vitce/3.2025.173

The EBAT architecture: An explainable blockchain for legal AI audits

Oleksii Shamov*

Intelligent Systems Researcher
Human Rights Educational Guild
18010, 40/28 Rizdviana Str., Cherkasy, Ukraine
http://orcid.org/0009-0009-5001-0526

Abstract. The integration of artificial intelligence into high-stakes domains like the justice system presents the "black box problem", where algorithmic opacity undermines fundamental legal principles and current blockchain-based auditing solutions fail to bridge the critical gap between a record's technical integrity and its value as interpretable legal evidence. This research aimed to develop and theoretically substantiate a novel audit system architecture that synergistically combines the cryptographic reliability of blockchain with the interpretive power of Explainable Artificial Intelligence (AI) to produce logs of AI decisions that are not only immutable but also legally significant and human-understandable. The methodology involved a systematic analysis and synthesis, including a review of publications from scientometric databases, an analysis of legal standards for digital evidence, and conceptual architectural design methods for information systems. The study proposed a new hybrid architecture, the "Explainable Blockchain Audit Trail", specifically designed to resolve this challenge. Its core novelty lies in a three-tiered structure that first mandates the generation of human-readable counterfactual explanations for every AI decision. Second, a complete and self-sufficient evidence package - containing the input data, model specifications, and the generated explanation - is securely stored in decentralised off-chain storage to ensure its integrity and availability. The third tier then creates an immutable "trust anchor" for this package on a permissioned blockchain, cryptographically linking all components and providing a permanent, tamper-proof record of the event. This comprehensive model ensures complete reproducibility of the decision-making process and establishes a robust, objective basis for judicial review and appeal. The proposed architecture provides a crucial theoretical foundation for developing practical tools for judges, lawyers, and regulators, ultimately aiming to enhance transparency and protect citizens' rights in the age of algorithmic decision-making by offering concrete mechanisms to challenge opaque conclusions

Keywords: legal tech; audit trail; digital evidence; immutable logs; justice

Introduction

The modern world is experiencing the fourth industrial revolution, driven by artificial intelligence (AI). AI technologies are penetrating the most conservative and responsible spheres of human activity, including finance, healthcare, and, most importantly, the justice system. The potential for optimising legal processes, analysing large arrays of case law, and assisting in decision-making is substantial. However, this technological expansion carries a systemic risk known as the "black box problem", where the internal logic of complex machine learning models remains hidden from human analysis. This opacity is not just a technical flaw but an existential threat to the rule of law, as it undermines principles of a fair trial and the right to a reasoned explanation. As highlighted in a review by S. Verma (2019) on the societal impact of algorithmic systems, such opaque models can perpetuate and even amplify existing biases, leading to discriminatory outcomes and deepening inequality, creating what are effectively "weapons of math destruction". Consequently, achieving "Trustworthy AI" is impossible without deeply integrated explainability mechanisms, a conclusion strongly supported by the comprehensive review from S. Ali *et al*. (2023), who argued that trust is unattainable until we can adequately understand and scrutinise AI-driven conclusions.

In response to these challenges, researchers have explored the use of blockchain technology to create immutable,

Suggested Citation:
Shamov, O. (2025). The EBAT architecture: An explainable blockchain for legal AI audits. *Information Technologies and Computer Engineering*, 22(3), 173-181. doi: 10.31649/vitce/3.2025.173

*Corresponding author

cryptographically protected audit trails for AI systems. A foundational review by K. Salah *et al*. (2019) outlined the significant potential of blockchain to enhance AI by providing a decentralised and tamper-proof mechanism for recording data and model actions, thereby addressing key issues of accountability and data integrity. The legal community is also actively examining these developments. For instance, X. Wang *et al*. (2024) explored the procedural implications of blockchain-based evidence in U.S. judicial processes, concluding that while blockchain's cryptographic security offers strong proof of integrity, its admissibility in court is still complex and requires clear standards for validation. This legal challenge is further emphasised by R. Bharati *et al*. (2024), whose work on digital evidence highlighted the significant hurdles that remain in ensuring its admissibility, given the ease of manipulation and the difficulty of establishing an unbroken chain of custody for digital artifacts. These issues are not merely academic; they are at the forefront of regulation, as analysed by S. Ramos & J. Ellul (2024), who detail how distributed ledger technology can be a critical tool for organisations seeking to comply with the stringent transparency and record-keeping requirements of the EU AI Act.

Despite these advancements, a critical aspect often remains underexplored: the very nature of what constitutes a valid explanation. The work of G. Vilone & L. Longo (2021) delved into the core notions of explainability, arguing that it is not a monolithic concept. They analysed various approaches to evaluating explanations, concluding that an effective explanation must be tailored to the context and the user, moving beyond purely technical metrics to ensure genuine human comprehension and utility.

A review of the current literature revealed a significant research gap. While substantial work has been done on using blockchain to ensure the technical integrity of AI logs, and a separate body of research has established the critical need for human-centric explainability, there is a lack of integrated frameworks that combine both elements to create legally admissible, interpretable evidence. Current models often focus on proving that a log has not been tampered with, but they do not address the equally important question of whether the decision recorded in that log is understandable and justifiable. This unexplored intersection between cryptographic immutability and semantic interpretability creates the urgent need for the present study.

Therefore, the purpose of this article was to develop and theoretically substantiate a new hybrid architecture for an AI decision audit system that synergistically integrates blockchain technology and Explainable AI (XAI) methods to create complete, immutable, and human-understandable evidence. To achieve this purpose, this article set out three primary objectives: first, to design a conceptual model of a hybrid architecture, the "Explainable Blockchain Audit Trail" (EBAT), which combines a permissioned blockchain, decentralised storage, and a mandatory explanation generation module; second, to analyse the potential of specific XAI methods, particularly counterfactual explanations, as the core tool for transforming technical logs into legally significant evidence within this architecture; and finally, to justify how the proposed architecture solves the "black box" problem in a legal context by meeting key requirements for digital evidence, including integrity, authenticity, and interpretability. The scientific novelty of this work lies in the development of the original three-tiered EBAT architecture, which synergistically combines three technologies (XAI, IPFS (InterPlanetary File System), Hyperledger Fabric) into a single system designed to meet the demands of the justice system by shifting the focus from simply recording a decision to recording its justification.

Literature Review

The problem of trust and accountability in artificial intelligence systems is multidisciplinary and is addressed in the works of many scholars in engineering, law, and social sciences. A review of the literature allowed for the identification of several key areas relevant to the topic of this study: the technical implementation of blockchain for data integrity, the legal standards for digital evidence, and the theoretical foundations of XAI. The first area concerns the technical aspects of using blockchain technology to ensure data integrity. One of the pioneering works in this domain was presented by A. Sutton & R. Samavi (2018), who proposed a conceptual framework for auditing AI systems using a tamper-proof log on a public blockchain. Their research established the foundational principle that cryptographic hashing of AI decisions and their inputs could provide a verifiable trail, though they also acknowledged the significant scalability and cost limitations inherent in using public ledgers like Bitcoin for high-frequency operations. Building on these early concepts, the review by K. Salah *et al*. (2019) provided a comprehensive map of the synergy between AI and blockchain. The authors systematically identified the main challenges and opportunities, concluding that while blockchain offers transformative potential for creating transparent and auditable AI systems, critical issues such as scalability, data privacy, and interoperability must be addressed through sophisticated architectural design. This idea was extended in broader research on blockchain-based trust management, as detailed in the survey by Y. Liu *et al*. (2023), which delved into the specific application of blockchain for trust management in the Internet of Things (IoT). Their work is highly relevant as it analysed how blockchain's decentralised and tamper-proof nature can establish verifiable trust between autonomous devices without a central authority, a principle directly applicable to creating trust in autonomous AI decisions in a legal setting. More recent works focused on specific regulatory contexts. For example, S. Ramos & J. Ellul (2024) directly connect distributed ledger technology to the stringent requirements of the European Union's AI Act. They argued that for high-risk AI systems, blockchain is not merely a technical option but a vital tool for demonstrating compliance, providing an

immutable record of the model's lifecycle, training data, and decision-making processes as required by the regulation. The practical application of these principles is now being actively implemented in various data-sensitive fields. M. Faruk *et al*. (2023) proposed a specific framework for securing electronic health records, utilising smart contracts and the InterPlanetary File System (IPFS) to ensure that patient data remains both confidential and integral, demonstrating a tangible use case for the hybrid on-chain/off-chain model. Similarly, Y. Zhang *et al*. (2023) explored the use of blockchain in federated learning, highlighting its role not only in preserving privacy but also in achieving verifiable fairness, a crucial concept for legal AI where algorithmic bias is a primary concern.

The second important area of research relates to legal standards and the admissibility of digital evidence. The work by X. Wang *et al*. (2024) provided a detailed analysis of how blockchain records might be treated within the U.S. judicial system. They concluded that while the cryptographic properties of blockchain provide a powerful technical argument for the authenticity and integrity of evidence, its legal admissibility is not automatic and hinges on procedural rules and the judiciary's evolving understanding of the technology. Their analysis underscored that a technical solution alone is insufficient without a clear legal framework for its acceptance in court. The principles for handling digital evidence, famously outlined by the UK's Association of Chief Police Officers (ACPO) and detailed in the work of E. Casey (2011), established the gold standard for digital forensics, requiring that no action should change original data and that a full audit trail of all processes must be maintained. This legal precedent set a high bar for any system claiming to produce court-admissible evidence. This point is further elaborated by R. Bharati *et al*. (2024), who examined the broader challenges of digital evidence in legal proceedings. They emphasised the inherent fragility of digital artifacts and the critical importance of maintaining a verifiable and unbroken "chain of custody", concluding that traditional methods are often inadequate and that new technological solutions are needed to meet longstanding evidentiary standards.

The third, and key for this work, area is Explainable AI, where the focus shifts from the integrity of the record to the intelligibility of its content. One of the foundational works that popularised model-agnostic explanations was the article by M. Ribeiro *et al*. (2016), which introduced the LIME (Local Interpretable Model-agnostic Explanations) technique. Their key insight was that one could explain any black-box model's prediction by approximating it with a simpler, interpretable model in the local vicinity of that prediction, a powerful concept that opened the door to a wide range of explanation methods. The scientific community has since conducted significant work systematising these approaches, as detailed in the comprehensive survey by R. Guidotti *et al*. (2018). They provided a robust taxonomy of XAI methods, classifying them into different categories and analysing the types of explanation problems they solve, which helps in selecting the appropriate technique for a given context. However, for the legal sphere, counterfactual explanations hold particular value. In their influential work, S. Wachter *et al*. (2018) argued that it is precisely counterfactual explanations ("the decision would have been different if…") that best meet the General Data Protection Regulation's (GDPR) requirements for a "right to explanation". They posited that such explanations are intuitively understandable and provide individuals with actionable recourse, a feature often missing from other explanation types. Building on this, the work of G. Vilone & L. Longo (2021) offered a crucial theoretical foundation by deconstructing the very concept of "explainability". They provided a detailed taxonomy of different types of explanations and, most importantly, argue that the evaluation of an explanation's quality cannot be purely technical. Instead, it must be human-centric, focusing on whether the explanation is genuinely useful, understandable, and satisfactory to the end-user, which strongly supports the need for legally-oriented explanation methods over generic technical outputs.

Despite significant progress in each of these distinct areas, their synergistic integration remains insufficiently explored. Most research combining AI and blockchain focuses on the technical mechanisms of immutability while overlooking the semantic content of what is being recorded. Conversely, works within XAI deeply analyse the nature of explanations but rarely propose robust, cryptographically secure frameworks for their storage and verification as legal evidence. It is this gap between proving a record is unchanged and proving it is understandable that the present study aimed to fill.

Materials and Methods

This study was theoretical and conceptual, and its methodology was based on a comprehensive approach that combines several analytical and design methods. The main stages of the work and the justification for the choice of methods were as follows: systematic literature review and analysis of the state of the problem; analysis of legal and technical standards; conceptual architectural design; synthesis and justification.

Stage 1. At the first stage, a systematic literature review was conducted to identify key concepts, existing solutions, and unresolved problems at the intersection of AI, blockchain, and jurisprudence. The search for sources was carried out in leading scientometric databases, Scopus and Web of Science. The following keywords and their combinations were used: "explainable AI", "blockchain audit trail", "AI accountability", "legal tech", "digital evidence", "immutable logs", 'Hyperledger Fabric governance", "counterfactual explanations". The inclusion criteria were: publications for 2018-2025, high citation index (for foundational works), relevance to the research topic, and presence of a DOI. The exclusion criteria were: purely marketing articles, non-peer-reviewed works, and overly narrow technical reports without analysis of the broader context. This method

allowed for the formation of a deep understanding of the current state of research, the identification of a scientific gap, and the justification of the work's relevance.

Stage 2. The second step was a detailed analysis of existing standards governing the handling of digital evidence. The key principles outlined in the ACPO good practice guide for digital evidence (2012) from the UK's Association of Chief Police Officers and the requirements for evidence authentication under the U.S. Federal Rules of Evidence (FRE 901) were analysed (The U.S. Congress, 2024). The purpose of this analysis was to define clear legal requirements that any AI audit system must meet for its results to be considered admissible evidence in court. This method allowed for the formulation of a set of criteria that became the basis for the architectural design.

Stage 3. This was the central stage of the research, where the method of conceptual modelling and architectural design of information systems was applied. Based on the requirements formulated in the previous stages, a new hybrid architecture, the Explainable Blockchain Audit Trail, was developed. The choice of specific technological components (Hyperledger Fabric, IPFS) was justified by their technical characteristics and suitability for corporate and legal sector tasks. Specifically, Hyperledger Fabric was chosen for its support of private transactions, high throughput, and modular architecture, which are advantages over public blockchains like Ethereum for this application (Androulaki *et al.*, 2018). IPFS was chosen for its content addressing mechanism, which guarantees the integrity of off-chain data (Benet, 2014). This method allowed for the creation of a detailed theoretical model of the system.

Stage 4. In the final stage, a synthesis of the obtained results was carried out. The proposed EBAT architecture was analysed for compliance with legal requirements and its ability to solve the identified "black box" problem. A comparative analysis of the EBAT architecture with existing approaches described in the literature was conducted to demonstrate its scientific novelty and potential advantages. The chosen methodology is valid for theoretical research as it allows for a systematic analysis of the problem, the formulation of a set of requirements, and, based on them, the development and justification of a new conceptual solution. The described sequence of steps is logical and can be reproduced by other researchers for verification or further development of the proposed ideas.

Results and Discussion

Based on the analysis and the applied methodology, a new conceptual architecture, the Explainable Blockchain Audit Trail, was developed. The EBAT architecture is a complex hybrid system that operates on three interconnected tiers. Each tier uses specific technologies to perform its part of the task, and together they create a single, cohesive, and legally significant audit trail (Fig. 1).

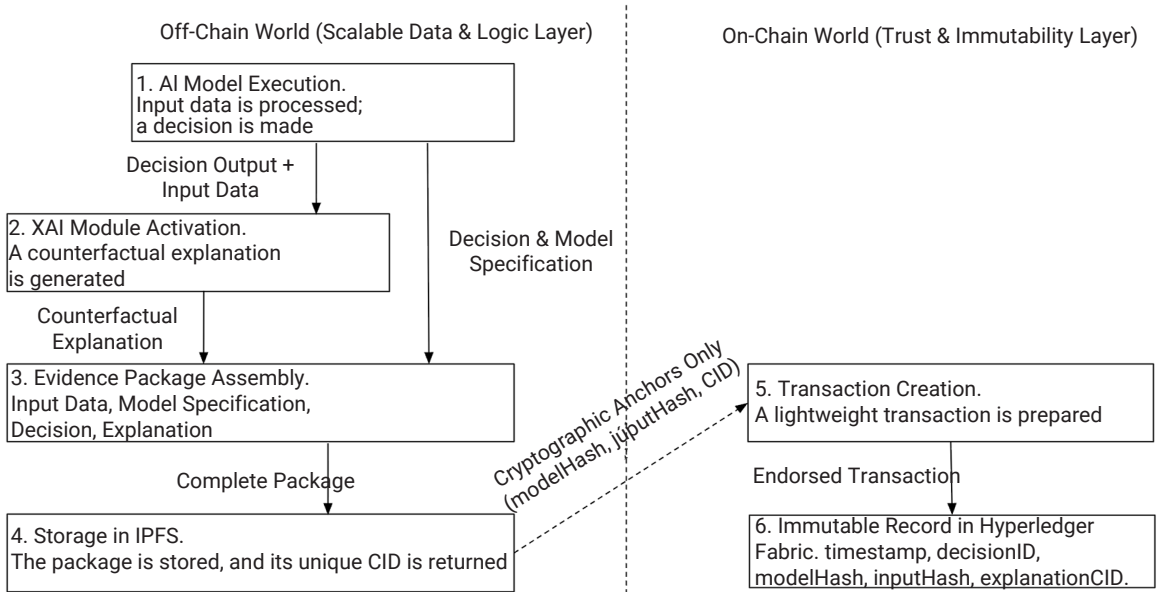


Figure 1. Overall architecture of the EBAT system

Source: completed by the author

Tier 1: Generation of the decision and the mandatory counterfactual explanation

This is the starting point and the ideological core of the entire architecture. The process begins when an AI system makes a decision with legal consequences (e.g., denying a loan, classifying evidence in a case, determining the risk of recidivism). In traditional systems, the logging process would end at this stage with the recording of technical parameters. In the EBAT architecture, however, this is just the beginning. Immediately after the decision is made, a mandatory integrated XAI module is activated This tier can be detailed into several sub-processes. First, the initial data (e.g., a loan applicant's form, digital evidence in a criminal case) passes through the main AI model. This model can

be anything from gradient boosting to a complex neural network but its conclusion (e.g., "deny", "approve", "high risk") is the trigger for the next step. Immediately after this conclusion is received, the same input data, along with the obtained result, is passed to the XAI module. The main requirement for this module is the generation of a counterfactual explanation. As justified in the work of S. Wachter *et al*. (2018), it is this type of explanation that is most valuable in the legal field, unlike other XAI methods such as SHAP (SHapley Additive exPlanations) or LIME.

While LIME (Ribeiro *et al*., 2016) explains which features were most important for a particular decision, a counterfactual explanation provides an actionable scenario. It answers the question: "What would need to be changed in the input data for the system to make an alternative, usually positive, decision?". This is a fundamental difference: instead of a passive statement ("you were denied because of low income"), the system provides an active instruction ("you would have been approved if your income was X higher"). For a lawyer, this means the ability to check whether the requirement for a "higher income" is discriminatory or unreasonable. For a citizen, it provides clear grounds for appeal.

Technically, generating counterfactual explanations is a complex optimisation task: the algorithm searches for minimal changes in the input feature vector that lead to a change in the model's classification while remaining plausible. Various algorithms exist for finding them, and the choice of a specific one depends on the type of AI model. It is important that the generation of such an explanation is not an option but a mandatory, atomic part of the decision-making process. Without the successful generation of an explanation, the entire process is considered incomplete and cannot be recorded. This ensures that no "silent" decisions without justification appear in the system.

Tier 2: Formation of the evidence package and its storage in decentralised storage

Since storing large volumes of data directly on the blockchain is technically inefficient and economically unfeasible, EBAT uses a hybrid model with the main part of the data stored off-chain. At this tier, the system automatically forms a single digital "evidence package". This package is a logically connected set of files that must be self-sufficient for the complete reproduction and analysis of the situation by an independent auditor.

Its structure can be represented in a standardised format, such as JSON or XML, including several key sections:

1. InputData. A section containing the complete input data used to make the decision. To maintain integrity and privacy, sensitive data can be hashed or pseudonymised, but their structure and values must be recorded.

2. ModelExecution. This section contains complete information about the AI model. This is not just a name, but the exact version (e.g., a Git commit), its unique identifier (e.g., a SHA-256 hash of the model's binary file), and a list of versions of key libraries (TensorFlow, PyTorch, scikit-learn) used during execution. This is absolutely critical for ensuring reproducibility, as even a minor change in a library version can affect the result.

3. DecisionOutcome. Contains both the concise conclusion of the system ("denied/approved") and possibly more detailed information, such as a confidence score (prediction probability).

4. Explanation. Here, the counterfactual explanation generated in Tier 1 is stored in full text format, as well as metadata about the generation process itself (which XAI algorithm was used, how long it took).

The formed evidence package, which can range from a few kilobytes to many megabytes, is uploaded to the IPFS. The choice of IPFS is fundamental. Unlike traditional centralised storage (e.g., Amazon S3), where the owner company can delete or change data, and access is via a variable link (URL), IPFS uses content addressing (Benet, 2014). The system computes a cryptographic hash of the entire package and uses this hash (CID – Content Identifier) as its unique and immutable address. This creates a powerful guarantee of integrity: any attempt to change even one byte in the evidence package (e.g., to forge an explanation) will result in a complete change of its CID. Thus, the link to this package that will be recorded on the blockchain will become invalid, instantly exposing the attempt to interfere. To guarantee constant data availability, the package must be "pinned" on several IPFS nodes, which prevents its accidental deletion.

Tier 3: Creation of the immutable "trust anchor" on a permissioned blockchain

This is the final tier, which ensures the cryptographic immutability, finality, and ordering of the entire process. After receiving the CID of the evidence package from IPFS, the system initiates a transaction on a permissioned blockchain. For this architecture, the use of the Hyperledger Fabric framework is proposed. As noted by developers and researchers, Fabric is ideally suited for corporate tasks due to its modular architecture, support for confidential channels (which allows different participants to see only relevant transactions), absence of a speculative cryptocurrency, and high performance compared to Proof-of-Work systems (Androulaki *et al*., 2018).

The transaction process in Fabric is multi-stage. A client application forms a transaction proposal and sends it for endorsement to endorsing nodes, which simulate the execution of the corresponding smart contract (chaincode). If the simulation is successful and the nodes reach an agreement, they sign the result and return it to the client. The client collects the endorsements and sends the final transaction to the ordering service, which guarantees a single order of transactions for the entire network. After this, transactions are grouped into blocks and sent to all nodes for validation and recording in their copies of the ledger.

The transaction recorded on the blockchain in the EBAT architecture is extremely lightweight and contains only cryptographic fingerprints, serving as a "trust

anchor". Its structure can be implemented as a call to a chaincode function, for example CreateAuditTrail, with the following arguments:

- timestamp: the exact timestamp of the event, provided by the ordering service;
- decisionID: a unique identifier for the decision, allowing it to be linked to other systems;
- modelHash: the hash of the model file (SHA-256), ensuring that the correct version of the model was used;
- inputHash: the hash of the input data to confirm their immutability;
- explanationCID: the key field the CID of the evidence package from IPFS, which is a cryptographic link to the full context.

Thus, an unbreakable, ordered chain of evidence is permanently recorded on the blockchain. Any attempt to change one of the elements (e.g., to claim that a different AI model was used) will be immediately detected, as the hash of the changed model will not match the one permanently recorded on the blockchain. This creates the highest level of trust and accountability necessary for legal practice.

As noted in the methodology, the EBAT architecture is not a purely technical solution; its design was purposefully shaped by fundamental legal standards governing the handling of digital evidence. An analysis of two key documents the UK's "ACPO good practice guide for digital evidence" (2012) and the U.S. Federal Rules of Evidence (The U.S. Congress, 2024) allowed for the formulation of a set of engineering criteria that any system aiming to create legally significant logs must meet. The "ACPO good practice guide for digital evidence" establishes four core principles that serve as the gold standard in digital forensics. Principle 1 states that no action should change data that may be relied upon in court. Principle 3 requires that a full audit trail of all processes applied to digital evidence be created and preserved, in such a way that a third party can repeat those processes and achieve the same result. Together, these principles form the requirement for technical integrity and reproducibility. On the other hand, the U.S. Federal Rules of Evidence, particularly Rule 901, require the authentication of evidence. This means the proponent must produce sufficient evidence to support a finding that the item is what the proponent claims it is. In a digital context, this translates to proving that a log file, screenshot, or other artifact is not a forgery and was created by the specific process and at the specific time alleged. Based on these legal norms, the following key criteria for the design of the EBAT architecture were formulated four criterions:

Criterion 1: guaranteed data immutability. The system must cryptographically guarantee that, once recorded, no component of the evidence package (input data, model, explanation) can be altered without detection.

Criterion 2: process integrity and reproducibility. The system must record not only the result but the entire chain of actions, and do so in enough detail that an independent auditor can fully reproduce and verify the decision-making process.

Criterion 3: evidentiary authentication. The system must create an irrefutable proof of the log's origin and time of creation, linking a specific decision to specific data, a specific model, and a specific timestamp.

Criterion 4: semantic clarity. The record must be not only technically sound but also understandable to non-technical participants in the legal process (judges, lawyers), meaning it must contain an interpretation, not just raw data.

The EBAT architecture was designed so that each of its tiers directly addresses these criteria. Compliance with Criterion 1 (Immutability) is ensured at Tiers 2 and 3. Storing the evidence package in IPFS guarantees that any change to the data will result in a change to its Content Identifier. Since this CID is permanently recorded on the blockchain at Tier 3, any discrepancy between the CID on the blockchain and the actual CID of an altered file will be instantly detected. This creates a dual layer of protection against tampering. Compliance with Criterion 2 (Integrity and Reproducibility) is achieved through the comprehensive structure of the "evidence package" at Tier 2. The inclusion of not only the input data but also the exact version of the AI model and its dependencies (libraries) is key to reproducibility. An independent auditor, possessing this package, can not only view the result but can re-run the exact same model on the exact same data to verify that the outcome is identical. Compliance with Criterion 3 (Authentication) is the primary function of Tier 3. The transaction in Hyperledger Fabric serves as a form of digital notarisation. It contains an accurate timestamp from the ordering service, cryptographic hashes of all key components (model, input data), and the package's CID. This record, endorsed by network participants and included in the immutable chain of blocks, constitutes powerful proof of authenticity that satisfies the requirements of FRE Rule 901. Compliance with Criterion 4 (Semantic Clarity) is the unique advantage realised at Tier 1. Unlike other systems, EBAT makes the generation of a counterfactual explanation a mandatory part of the process. By including this human-readable explanation in the evidence package, the architecture transforms the "black box" into a transparent process, the results of which can be meaningfully analysed in court, not just technically verified. Thus, the EBAT architecture does not merely use blockchain as a technology for recording hashes; it is a holistic system designed "from law to code", where every architectural choice is justified by the specific requirements of legal evidentiary standards.

The proposed EBAT architecture is a direct response to the gaps identified in the literature review, offering a synthesised solution where existing research often focuses on separate components of the problem. Its effectiveness can be best understood by comparing it to the current scientific discourse in both the blockchain and Explainable AI domains. The foundation of the EBAT architecture rests on the principles of blockchain technology, which are extensively covered in the literature. The work of Y. Yuan & F. Wang (2019) provided a broad overview of blockchain

models and applications, establishing the technology as a robust framework for creating decentralised trust and ensuring data integrity through cryptographic linkage. The architecture builds upon this general model, but in a highly specialised manner. EBAT applied these principles to the niche but critical task of creating legally admissible evidence. This aligns with the vision outlined by K. Salah *et al*. (2019), who identified the potential for blockchain to bring accountability to AI but also highlighted the challenges of scalability and privacy. The EBAT architecture directly addresses these challenges through its hybrid on-chain/off-chain design, using Hyperledger Fabric for efficient, permissioned transactions and IPFS for scalable off-chain storage. This design choice is further validated by recent specialised applications. For instance, the auditing scheme for educational data proposed by F. Yu *et al*. (2024) demonstrated a modern blockchain application designed for trusted data detection. However, their model, while effective for verifying data integrity, exemplifies the very gap EBAT aims to fill: it ensures that the educational record is untampered but does not provide any mechanism to explain why an AI might have made a certain assessment based on that data. EBAT, in contrast, considers the integrity of the record and the intelligibility of its content to be equally important. This brings to the second pillar of our architecture: Explainable AI.

The fundamental challenge that XAI seeks to address is thoroughly documented in the comprehensive survey by A. Adadi & M. Berrada (2018). They provided a detailed taxonomy of the "black box" problem across various AI models and survey the landscape of explanation techniques designed to "peek inside". Their work clarified that there is no one-size-fits-all solution for explainability; the choice of method is highly context-dependent. The EBAT architecture acknowledges this by deliberately selecting a specific type of explanation-counterfactuals-based on their unique suitability for the legal domain. This choice is strongly supported by the legal and ethical analysis of S. Wachter *et al*. (2018), who argued that counterfactuals ("the outcome would have been different if…") provided actionable recourse and directly align with the principles of the GDPR's "right to explanation". While other techniques might provide technical insights, counterfactuals offer a narrative that is intuitively understandable to judges, lawyers, and the individuals affected by a decision. This aligns with the human-centric perspective advocated by G. Vilone & L. Longo (2021), who argued that the quality of an explanation should be judged not by its technical elegance but by its usefulness and comprehensibility to the end-user.

Thus, the primary contribution of the EBAT architecture to the scientific discourse is its synergistic synthesis. It moves beyond the siloed approaches prevalent in the literature. Unlike blockchain frameworks that focus solely on data integrity (like the one proposed by F. Yu *et al*. (2024)), and unlike theoretical XAI models that lack a secure, immutable storage mechanism for the explanations they generate (as is common in the works surveyed by A. Adadi & M. Berrada (2018)), EBAT binds the explanation to the record in a cryptographically inseparable manner. It transforms the blockchain from a simple timestamping service for opaque data into a permanent, verifiable ledger of justified decisions. By doing so, it provides a tangible engineering solution that addresses the complex socio-technical problem of building trust in AI within high-stakes, adversarial environments like the justice system.

Conclusions

This study conducted a comprehensive analysis of the accountability problem of artificial intelligence systems within the legal context, identifying a key gap between existing technical solutions that ensure the immutability of logs via blockchain and the fundamental legal requirement for their interpretation. In response to this challenge, a new hybrid architecture, the Explainable Blockchain Audit Trail, was developed and theoretically substantiated. The work provided a detailed description of its three-tiered structure, which synergistically combines Explainable AI methods, decentralised storage, and a permissioned blockchain with the goal of creating technically robust and legally significant evidence that complies with modern legal standards.

The conducted research allowed to draw several important conclusions. Firstly, the rapid implementation of artificial intelligence systems in the justice system and other regulated fields creates an acute need for mechanisms that ensure their transparency, accountability, and trust. Existing logging approaches, even with the use of blockchain technology, are insufficient as they only guarantee the technical immutability of records but do not solve the fundamental "black box" problem, leaving the logic of decisions opaque to lawyers and citizens. Secondly, for an audit trail to have real evidentiary value in court, it must meet not only technical criteria for integrity but also legal requirements for interpretation and comprehensibility. This means that the system must record not only the fact of a decision but also its justification in a human-accessible form.

The proposed architecture EBAT is a comprehensive solution that eliminates the identified gap between the technical and legal components of an audit. Through the synergistic integration of three key components a module for generating counterfactual explanations, the decentralised storage IPFS, and the permissioned blockchain Hyperledger Fabric-the EBAT architecture creates an audit trail that is simultaneously immutable, complete, reproducible, and, most importantly, legally significant. This allows for a transition from passive data recording to an active accountability system where every AI decision can be effectively verified and challenged. Thus, the research has achieved its goal by proposing an innovative engineering concept that has significant potential for building trust in automated systems in critically important areas.

Prospects for further research are primarily focused on the practical implementation and experimental validation of the proposed EBAT architecture. Creating a software prototype and testing it in realistic scenarios, such as credit

scoring or legal document analysis, is a necessary next step to assess its real-world performance, security, and cost-effectiveness. The results of such empirical validation would, in turn, provide the crucial foundation for the second priority area: the development of industry standards for the format and content of legally significant explanations of AI decisions, which would facilitate their unification and interoperability across different systems and jurisdictions.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138-52160. doi: 10.1109/ACCESS.2018.2870052.

[2] Androulaki, E., *et al*. (2018). Hyperledger fabric: A distributed operating system for permissioned blockchains. In *Proceedings of the thirteenth EuroSys conference* (pp. 1-15). New York: ACM. doi: 10.1145/3190508.3190538.

[3] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, article number 101805. doi: 10.1016/j.inffus.2023.101805.

[4] ACPO good practice guide for digital evidence (Version 5). (2023). Retrieved from https://www.digital-detective.net/digital-forensics-documents/ACPO_Good_Practice_Guide_for_Digital_Evidence_v5.pdf.

[5] Benet, J. (2014). IPFS - content addressed, versioned, P2P file system. *ArXiv*. doi: 10.48550/arXiv.1407.3561.

[6] Bharati, R., Khodke, P., Khadilkar, C., & Bawiskar, S. (2024). Forensic bytes: Admissibility and challenges of digital evidence in legal proceedings. *International Journal of Scientific Research in Science and Technology*, 11(16), 24-35. doi: 10.2139/ssrn.4896874.

[7] Casey, E. (Ed.) (2011). *Digital evidence and computer crime: Forensic science, computers, and the internet* (3rd ed.). Amsterdam: Academic Press.

[8] Faruk, M., Shahriar, H., Saha, B., & Barek, A. (2023). Security in electronic health records system: Blockchain-based framework to protect data integrity. In Y. Maleh, M. Alazab & I. Romdhani (Eds.), *Blockchain for cybersecurity in cyber-physical systems. Advances in information security* (Vol. 102, pp. 125-137). Cham: Springer. doi: 10.1007/978-3-031-25506-9_7.

[9] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5), 1-42. doi: 10.1145/3236009.

[10] Liu, Y., Wang, J., Yan, Z., Wan, Z., & Jäntti, R. (2023). A survey on blockchain-based trust management for Internet of Things. *IEEE Internet of Things Journal*, 10(7), 5898-5922. doi: 10.1109/JIOT.2023.3237893.

[11] Ramos, S., & Ellul, J. (2024). Blockchain for Artificial Intelligence (AI): Enhancing compliance with the EU AI Act through distributed ledger technology. A cybersecurity perspective. *International Cybersecurity Law Review*, 5, 1-20. doi: 10.1365/s43439-023-00107-9.

[12] Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). New York: ACM. doi: 10.1145/2939672.2939778.

[13] Salah, K., Rehman, M., Nizamuddin, N., & Al-Fuqaha, A. (2019). Blockchain for AI: Review and open research challenges. *IEEE Access*, 7, 10127-10149. doi: 10.1109/ACCESS.2018.2890507.

[14] Sutton, A., & Samavi, R. (2018). Tamper-proof privacy auditing for artificial intelligence systems. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence (IJCAI-18)* (pp. 5374-5378). Stockholm: IJCAI. doi: 10.24963/ijcai.2018/756.

[15] The U.S. Congress. (2024). *Federal rules of evidence*. Retrieved from https://www.uscourts.gov/file/78325/download.

[16] Verma, S. (2019). Weapons of math destruction: How big data increases inequality and threatens democracy. *The Journal for Decision Makers*, 44(2), 97-98. doi: 10.1177/0256090919853933.

[17] Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89-106. doi: 10.1016/j.inffus.2021.05.009.

[18] Wang, X., Wu, Y.C., & Ma, Z. (2024). Blockchain in the courtroom: Exploring its evidentiary significance and procedural implications in U.S. judicial processes. *Frontiers in Blockchain*, 7, article number 1306058. doi: 10.3389/fbloc.2024.1306058.

[19] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *ArXiv*. doi: 10.48550/arXiv.1711.00399.

[20] Yuan, Y., & Wang, F. (2019). Blockchain and cryptocurrencies: Model, techniques, and applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(9), 1421-1428. doi: 10.1109/TSMC.2018.2854904.

[21] Yu, F., Lu, Q., Meng, L., Peng, J., Xi, J., & Li, X. (2024). A blockchain-based auditing scheme for educational data supporting trusted detection. In *Twelfth international conference on advanced cloud and big data (CBD)* (pp. 184-189). Brisbane: IEEE. doi: 10.1109/CBD65573.2024.00042.

[22] Zhang, Y., Tang, Y., Zhang, Z., Li, M., Li, Z., Khan, S., Chen, H., & Cheng, G. (2023). Blockchain-based practical and privacy-preserving federated learning with verifiable fairness. *Mathematics*, 11(5), article number 1091. doi: 10.3390/math11051091.

Архітектура EBAT: пояснюваний блокчейн для юридичного аудиту ШІ

Олексій Шамов

Дослідник інтелектуальних систем
Громадська організація «Освітня гільдія прав людини»
18010, вул. Різдвяна, 40/28, м. Черкаси, Україна
http://orcid.org/0009-0009-5001-0526

Анотація. Інтеграція штучного інтелекту у сфери з високим рівнем відповідальності, як-от правосуддя, створює «проблему чорної скриньки», де непрозорість алгоритмів підриває фундаментальні правові принципи, а наявні рішення для аудиту на основі блокчейну не здатні подолати критичний розрив між технічною цілісністю запису та його цінністю як юридичного доказу, що піддається інтерпретації. Це дослідження мало на меті розробити та теоретично обґрунтувати нову архітектуру системи аудиту, яка синергетично поєднує криптографічну надійність блокчейну з інтерпретаційною потужністю пояснюваного штучного інтелекту (ШІ) для створення логів рішень, що є не лише незмінними, але й юридично значущими та зрозумілими для людини. Методологія включала системний аналіз та синтез, огляд публікацій з наукометричних баз даних, аналіз правових стандартів для цифрових доказів та методи концептуального архітектурного проектування інформаційних систем. У дослідженні запропонована нова гібридна архітектура «Explainable Blockchain Audit Trail», спеціально розроблену для розв'язання цієї проблеми. Її новизна полягає у трирівневій структурі, яка, по-перше, вимагає обов'язкової генерації зрозумілих для людини контрфактичних пояснень для кожного рішення ШІ. По-друге, повний та самодостатній пакет доказів, що містить вхідні дані, специфікації моделі та згенероване пояснення, надійно зберігається у децентралізованому off-chain сховищі для гарантування його цілісності та доступності. Третій рівень створює незмінний «якір довіри» для цього пакету у приватному блокчейні, криптографічно пов'язуючи всі компоненти та забезпечуючи постійний, захищений від втручання запис про подію. Ця комплексна модель забезпечує повну відтворюваність процесу прийняття рішень та створює надійну, об'єктивну основу для судового перегляду та апеляції. Запропонована архітектура надає ключову теоретичну основу для розробки практичних інструментів для суддів, адвокатів та регуляторів, кінцевою метою якої є підвищення прозорості та захист прав громадян в епоху алгоритмічного прийняття рішень шляхом надання конкретних механізмів для оскарження непрозорих висновків

Ключові слова: юридичні технології; аудиторський слід; цифрові докази; незмінність логів; правосуддя

Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 01.08.2026 Revised: 13.11.2025 Accepted: 23.12.2025

UDC 657.37:622 DOI: 10.31649/vitce/3.2025.182

Use of intelligent algorithms in virtual healthcare computer systems: From diagnosis to personalised treatment

Mykola Khrulov*

Postgraduate Student
Cherkasy State Technological University
18006, 460 Shevchenko Blvd., Cherkasy, Ukraine
https://orcid.org/0000-0001-8532-0967

Tetiana Myroniuk

PhD in Technical Sciences, Associate Professor
Cherkasy State Technological University
18006, 460 Shevchenko Blvd., Cherkasy, Ukraine
https://orcid.org/0000-0002-7588-1055

Abstract. The study aimed to theoretically substantiate approaches to the effective implementation of intelligent algorithms in virtual medicine. The methodology was based on theoretical, analytical, and normative-prognostic analysis of the effectiveness and development of intelligent technologies in digital healthcare. The study established that artificial intelligence (AI) is transforming approaches to the collection, analysis and use of medical data. Virtual medicine uses machine learning for diagnosis, prediction and personalised treatment, increasing the accuracy of decisions and reducing the burden on doctors. Machine learning methods are effective for processing electronic medical records and laboratory data, while deep learning forms the basis of virtual medicine by automating the analysis of large amounts of information. Generative models create synthetic medical data and clinical scenarios, supporting the development of personalised medicine and the concept of "digital twins". Multimodal systems combine different types of data, providing a comprehensive analysis of the patient's condition and more accurate clinical predictions. The benefits of AI implementation included an 18-25% increase in diagnostic accuracy, a 20-30% reduction in working hours among doctors, expanded access to medicine in remote regions, and lower healthcare costs. The main risks are issues of data security, explainability, ethics, bias, and doctor trust, which necessitate transparency, control, and legal regulation. The European Union has specific legislation that sets requirements for the safety and transparency of medical AI systems, while Ukraine's regulatory framework is still in the process of being developed. To improve virtual medicine, it is advisable to implement explainable AI, integrate Large Language Models with data protection, apply federated learning, generative simulations and blockchain following ethical and legal standards. The results of the study can be used by specialists when making decisions on the selection and application of intelligent algorithms in medical institutions, research centres, and the IT sphere of healthcare

Keywords: multimodal medical data analytics; digital monitoring; generative models for simulations; explainability of clinical decisions; telemedicine; security and privacy of medical data

Introduction

The rapid development of information technology and the growth of digital medical data volumes necessitate the introduction of intelligent algorithms into healthcare computer systems. The medicine of the 21st century increasingly requires effective tools capable of analysing large amounts of clinical, genomic and behavioural information, identifying hidden patterns and providing personalised recommendations. In this context, the integration of Machine Learning (ML) and Deep Learning (DL) methods into virtual medical systems forms the basis of a new stage of digital transformation in the industry, covering diagnosis, prognosis, treatment and real-time patient monitoring.

Suggested Citation:
Khrulov, M., & Myroniuk, T. (2025). Use of intelligent algorithms in virtual healthcare computer systems: From diagnosis to personalised treatment. *Information Technologies and Computer Engineering*, 22(3), 182-194. doi: 10.31649/vitce/3.2025.182

*Corresponding author

Intelligent algorithms are considered the basis of clinical diagnostics, which can be used for the analysis of large arrays of clinical data, the identification of latent patterns, and the optimisation of the physician's decision-making process. S. Mizna *et al.* (2025) proved that the use of ML models in clinical decision support systems significantly reduces the number of diagnostic errors. The study demonstrated that a combination of structured and unstructured medical data increases the accuracy of automated systems by more than 20%, indicating the basic effectiveness of artificial intelligence (AI) in virtual medicine. S.R. Abbas *et al.* (2025) demonstrated that smart healthcare systems based on deep neural networks effectively use biomedical signals and visual data for early detection of pathological changes. This confirmed that AI algorithms can not only improve diagnostic accuracy but also support doctors in making decisions in complex clinical scenarios.

A significant portion of studies is devoted to the personalisation of treatment, where intelligent systems adapt therapeutic decisions to the individual characteristics of the patient. X. Guo & Y. Li (2024) demonstrated that health information systems form the basis for integrating heterogeneous data, from medical records to genetic profiles, which creates opportunities for predicting the course of diseases. Similar conclusions were made by H.B. Clark *et al.* (2024), emphasising the need to create hybrid models that combine statistical methods and deep neural networks to predict treatment outcomes. This has laid the groundwork for the development of personalised medicine, where AI algorithms act not as an auxiliary tool but as an active element in clinical decision-making, facilitating the transition from universal therapeutic approaches to individually tailored treatment strategies.

The issues of security and ethics in the use of AI in medicine remain central. M.M. Khan *et al.* (2025) determined that the lack of transparent data management policies threatens trust in AI decisions. The authors formulated a list of technical and organisational measures necessary to create a safe environment for the use of AI tools in clinics, which laid the foundation for the concept of "trusted AI in healthcare". Explainable AI is also one of the critical areas of development for medical AI. R. Alkhanbouli *et al.* (2025) systematised approaches to building transparent AI models in medicine and showed that the use of Explainable AI reduces the risk of misinterpretation of results by doctors. The study proved that explainable algorithms increase the trust and clinical acceptability of systems, forming the theoretical basis for their implementation in healthcare.

Ukrainian researchers are also contributing to the scientific discourse on the implementation of intelligent technologies in healthcare. O. Boychenko & T. Bubliy (2024) highlighted the potential of using AI algorithms to optimise diagnostic processes in the national healthcare system, while emphasising the limitations associated with the lack of high-quality medical databases and standardised information processing protocols. V.O. Korotka & V.A. Mokrynskyi (2024) highlighted the technical and organisational barriers to the digitalisation of Ukrainian medicine, in particular the shortage of qualified personnel capable of interpreting the results generated by DL algorithms. N. Sofilkanych *et al.* (2023) determined that the Ukrainian medical sector is only beginning to systematically implement intelligent solutions, but the prospects for their application (from personalised therapy to telemedicine services) are significant and strategic for the country. The overall contribution of Ukrainian researchers lies in creating a conceptual framework for the development of national digital medicine, which combines the technical, organisational and managerial aspects of AI implementation. This facilitates the adaptation of global approaches to the local context and lays the foundation for further interdisciplinary research in the field of virtual healthcare.

Despite the availability of scientific publications, several unresolved issues remain. Most studies emphasise technical aspects, neglecting issues of clinical validation, explainability of decisions, and user trust. There are no uniform standards for integrating intelligent algorithms into virtual healthcare systems that can ensure the compatibility of different platforms and the protection of confidential data. The mechanisms of personalising treatment based on a comprehensive analysis of multimodal medical data (images, texts, biosignals, genomic profiles) remain insufficiently studied. Therefore, the study aimed to provide a theoretical justification for the use of AI to improve the analytical, diagnostic, and prognostic capabilities of virtual healthcare systems. To achieve this goal, the following tasks were set: to analyse approaches to the application of AI in diagnosis and treatment, to identify the advantages and risks of using intelligent algorithms to assess the potential of personalised treatment models, and to formulate recommendations for the safe integration of AI technologies into virtual healthcare systems, in particular in Ukraine.

Materials and Methods

The study was analytical and theoretical in nature and was based on the systematisation, generalisation and critical analysis of scientific sources covering the application of AI in virtual medicine. The source base consisted of 27 scientific publications from 2022 indexed in the Scopus and Web of Science databases, an analytical report by the European Commission (2024), the Artificial Intelligence Act (2024) and data from the Ministry of Digital Transformation of Ukraine (2023) in the field of regulation, which define the principles of security, transparency and certification of medical AI systems. The sources were selected based on criteria of scientific reliability, relevance and representativeness for modern trends in AI development in medicine as of 2025. The research procedure involved the phased application of theoretical methods – analysis and comparison of ML concepts, DL, generative AI and multimodal architectures, summarising the results of practical AI implementations in clinical practice, comparative analysis of legal regulation in the EU and Ukraine, analytical synthesis

of conclusions with the definition of strategic directions for the development of intelligent medicine.

Using methods of theoretical generalisation and comparative analysis, based on data from scientific sources, intellectual technologies in the field of virtual medicine were analysed: ML (Support Vector Machine (SVM) and Decision Trees models), DL (Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and Transformers), generative AI (Generative Adversarial Networks (GAN), Diffusion Models, Large Language Models (LLM) (in particular Generative Pre-trained Transformer (GPT), Medical Patient Language Model (MedPaLM), Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT)) and multimodal architectures. The selection of these algorithms was based on the principle of representativeness in terms of key AI areas and their practical significance for medicine: ML for interpreted processing of structured data, DL for high-precision analysis of images, signals and texts, generative AI for synthesising medical data, and multimodal systems for integrating different types of information into a single clinical context. To evaluate the effectiveness of each area, a five-point rating scale was used based on the criteria of interpretability, accuracy, data requirements, flexibility, and practical applicability in medicine, which reflect the key requirements for AI algorithms in a clinical environment: clarity of results, reliability of predictions, resource efficiency, adaptability to different types of data, and real usefulness for medical decisions. The task of this stage was to systematically summarise the capabilities of different classes of intelligent algorithms for improving the efficiency, accuracy and adaptability of virtual medicine to create a theoretical basis for developing a generalised model for their implementation in digital medical systems.

To analyse practical areas of AI application, a classification method was used to identify key areas for integrating intelligent algorithms into medical practice: intelligent diagnostics, patient monitoring, personalised treatment, and telemedicine with clinical decision support systems. The selection of these areas is justified by their systemic importance in the structure of medical care: intelligent diagnostics determines the accuracy of clinical decisions, monitoring ensures timely response, personalised treatment increases the effectiveness of therapy, and telemedicine expands the availability of services. The goal of this stage was to create an analytical framework describing the real directions of AI integration into medical practice. The effectiveness of AI in virtual medicine was assessed according to three groups of indicators: clinical results (diagnostic accuracy, data processing speed), economic effects (cost reduction, productivity improvement), and social consequences (improved accessibility of medical services). The selection of areas was based on criteria of practical significance, measurability of effects, and socio-economic effectiveness. The task of this stage was not only to quantitatively and qualitatively assess the effectiveness of the application of intelligent technologies in key processes of virtual medicine, but also to justify their socio-economic feasibility as a basis for further optimisation of implementation models in healthcare systems.

Critical analysis and a comparative-normative approach were used to identify ethical, safety, and legal risks and barriers to the implementation of intelligent systems in medical practice. The EU regulatory framework was compared based on the provisions of the European Commission (2024) and the Artificial Intelligence Act (2024), which define the principles of safety, transparency and certification of medical AI systems, and Ukraine's Ministry of Digital Transformation of Ukraine (2023) document, which outlines the stages of forming a national regulatory system in the field of AI. This helped to identify and classify the main barriers to the safe integration of intelligent systems into clinical practice, as well as to define directions for improving the ethical, legal, and security mechanisms governing their use.

Results and Discussion

Technological and applied aspects of using intelligent algorithms in virtual medicine

Intelligent technologies have changed the way medical data is collected and analysed. Virtual medicine, including telemedicine and remote monitoring, uses ML and DL methods to automate diagnosis and prognosis. Algorithms process large amounts of information, identifying patterns, forming risk profiles, and proposing personalised treatment plans. This increases diagnostic accuracy and reduces the burden on medical staff, optimising clinical decision-making. AI methods – ML, DL, generative AI, and multimodal architectures – differ in their data processing principles, interpretability, accuracy, and clinical application capabilities. A comparative overview of these technologies is presented in Table 1.

Table 1. Comparative evaluation of AI methods in virtual medicine

Direction	Key algorithms	Data types/ processing	Main areas of medical application	Benefits	Limitations/ challenges
ML	SVM, Random Forest, K-Nearest Neighbours, Decision Trees	Tabular, laboratory, clinical data	Classification, risk prediction, personalised medicine	High transparency, ease of learning, stability with small samples	Limited accuracy on complex, unlabelled data, poor scalability
DL	CNN, RNN, LSTM, Transformers	Medical images, time series, signals (EEG, ECG)	Computer diagnostics, analysis of visual and biomedical signals	High accuracy, automatic feature extraction, noise resistance	Low interpretability, need for large amounts of data, risk of overfitting

Table 1. Continued

Direction	Key algorithms	Data types/ processing	Main areas of medical application	Benefits	Limitations/ challenges
Generative AI	GAN, Diffusion Models, LLM (GPT, MedPaLM, BioBERT)	Text, synthetic and multimodal data	Data generation, telemedicine, clinical forecasting, and reporting	Flexibility, ability to create synthetic samples, support for clinical decisions	Ethical risks, complexity of validation, and high energy consumption
Multimodal architectures	HAIM, DRAGONET, MOICVAE	Integration of visual, textual, genomic, and tabular data	Comprehensive diagnostics, personalised treatment, digital biomarkers	Highest accuracy and completeness of data, improved consistency of results	High resource requirements, complex interpretation

Note: HAIM (H – human, A – algorithm, I – information, M – machine), MOICVAE – multi-omics informed conditional variational autoencoder

Source: compiled by the authors based on L. Soenksen *et al*. (2022), M. Cascella *et al.* (2023), W. Abbaoui et al. (2024), H. Nilius *et al*. (2024), J. Pool *et al*. (2024), S. Thapa *et al*. (2024)

ML remains the basic tool for processing structured medical data, such as electronic health records (EHR), laboratory indicators, or disease progression statistics. SVM, Random Forest, K-Nearest Neighbours, and Decision Trees algorithms provide high explainability of decisions and accuracy with small samples. They are used to classify tumour types, assess the risk of hospitalisation, predict complications, etc. The main advantage of ML is interpretability and low computational complexity, but its effectiveness decreases in cases of unstructured data. DL has become the core of virtual medicine due to its ability to automatically identify patterns in large data sets. CNN, RNN, LSTM, and Transformer architectures achieve high accuracy in the analysis of medical images, time series, and clinical texts. CNNs enable the recognition of pathologies in computed tomography, magnetic resonance imaging, and X-rays, while RNNs and LSTMs work effectively with physiological signals (ECG, EEG). Transformers form the basis of intelligent clinical assistants, triage systems, and decision support. The high accuracy of DL is combined with low interpretability and high requirements for data quality and computing resources.

Generative architectures (GAN, Diffusion Models, LLM) create a new level of adaptability and personalisation. GANs are used to create synthetic medical images, expand datasets, and simulate clinical scenarios, while LLMs (GPT, MedPaLM, BioBERT) are used to process clinical texts, form conclusions, and interact with doctors and patients. They increase the accuracy of predictions and can be used for the creation of digital twins of patients, but require mechanisms for ethical monitoring and validation of results. ML provides transparency and speed of analysis, DL provides the highest accuracy and generalisation ability, and generative AI provides flexibility and innovation in personalised medicine. Their comprehensive combination creates the basis for the formation of an adaptive, explainable, and secure virtual medical ecosystem. Additionally, Table 2 presents an assessment of AI methods according to key performance criteria.

Table 2. Systematic assessment of AI applications in virtual healthcare

Method	Interpretability	Accuracy	Data requirements	Flexibility	Use in medicine
ML	5	3	2	3	4
DL	2	5	5	4	5
Generative AI	2	4	5	5	3
Multimodal architectures	3	5	5	4	5

Source: compiled by the authors based on L. Soenksen *et al.* (2022), M. Cascella *et al.* (2023), W. Abbaoui *et al*. (2024), H. Nilius *et al*. (2024), J. Pool *et al*. (2024), S. Thapa *et al*. (2024)

ML maintains the highest interpretability and stability on small samples, but is inferior to DL systems in terms of accuracy. The latter demonstrate the best results in visual diagnostics and signal analysis, but are highly dependent on data volume and resources. Generative models provide flexibility and the ability to create synthetic medical data, but require proven validation mechanisms and ethical controls. Multimodal architectures combine the advantages of all approaches – high accuracy, generalisation ability, and practical applicability, forming the basis for integrated virtual medicine systems. The differences between the approaches indicate the need for the comprehensive use of different technologies within a single healthcare system.

As of 2025, one of the areas of development of intelligent technologies in medicine is multimodal systems that combine different types of data (visual, textual, signal), as well as demographic and behavioural characteristics of the patient. Such systems can be used to create comprehensive clinical profiles that provide a more accurate determination of health status, improve the quality of diagnosis and the effectiveness of personalised treatment. Thanks to their ability to integrate heterogeneous sources of information, multimodal architectures are becoming the basis of a new paradigm of analytics in virtual healthcare. A key component of such systems is the cross-modal attention mechanism, which can be used in the model to coordinate

information between different modalities, including images, numerical data, and medical report texts. This means that the system can simultaneously analyse magnetic resonance imaging results, laboratory indicators, electrocardiograms and clinical records to form an integrated assessment of the patient's condition. Such approaches provide synergy between different data sources, which increases the accuracy of clinical predictions and can be used for the identification of complex inter-system patterns (Abbaoui *et al.*, 2024; Nilius *et al.*, 2024).

The combination of medical imaging results, laboratory tests and text-based medical records can be used for effective prediction of the risks of recurrent stroke, complications from diabetes mellitus or other chronic diseases. Thanks to these technologies, virtual healthcare systems are gradually transforming into intelligent clinical assistants capable of analysing large amounts of diverse information in real time. This approach can be used to create dynamic, context-sensitive medical platforms that support clinical decision-making and promote the development of fully-fledged personalised medicine (Abbaoui *et al.*, 2024). Table 3 summarises the practical applications of ML and DL technologies in various components of digital healthcare.

Table 3. The use of intelligent algorithms in virtual healthcare systems

Direction	Characteristic	Implementation
Intelligent diagnostics based on deep learning	CNN, ViT, U-Net for medical image analysis (X-ray, computed tomography, magnetic resonance imaging, ultrasound), multimodal systems improve diagnostic accuracy and reliability	HAIM framework – image and text integration, diagnostic accuracy +6-33%, cGAN – synthetic data for oncology cases, image quality improvement, U-HPNet, GP-GAN – progression prediction for nodes and glioblastoma
Patient monitoring and digital biomarkers	RNN, LSTM, and Transformers algorithms for time series analysis; AI systems create digital biomarkers for predicting chronic diseases	GluGAN – glucose data generation, monitoring accuracy +15%, ML models integrate genomic and mass spectrometry data for patient phenotyping, use of wearables for early detection of complications
Personalised treatment and recommendation systems	Personalisation of therapy based on EHR, genomic and behavioural data, and the generation of digital twins	MOICVAE – drug sensitivity prediction in cancer (GDSC, CCLE), DRAGONET – generation of new drug candidates, generative AI for synthesising personalised treatment scenarios
Telemedicine and clinical decision support systems	LLM (ChatGPT, MedPaLM) for text analysis, consultations and report generation, hybrid models for decision support	ChatGPT, MedPaLM – asynchronous consultations and automatic reports, cGAN + Random Forest – forecasting the volume of teleconsultations, NLP assistants for reminders and medication planning

Note: U-HPNet – U-Net-based hierarchical prediction network; GP-GAN – generative prediction generative adversarial network; GDSC – genomics of drug sensitivity in cancer; CCLE – cancer cell line encyclopedia
Source: compiled by the authors based on L. Soenksen *et al.* (2022), M. Cascella *et al*. (2023), W. Abbaoui *et al*. (2024), I. Ghebrehiwet *et al*. (2024), H. Nilius *et al*. (2024), J. Pool *et al*. (2024), S. Thapa *et al.* (2024), E. Kumah (2025)

The application of intelligent algorithms in virtual medicine covers diagnostics, monitoring, personalisation of treatment and telemedicine. In particular, the HAIM framework (a conceptual approach used to integrate and manage various technologies and processes in medical and technological systems: H – human, A – algorithm, I – information, M – machine) has demonstrated a 6-33% increase in the accuracy of chest pathology diagnosis by integrating images, text, and time series (Ministry of Digital Transformation of Ukraine, 2023). cGAN (Conditional Generative Adversarial Network) architectures are successfully used to generate synthetic medical images in telemedicine for oncology, which improves the quality of data from portable devices. In the field of monitoring, the GluGAN model (a specific variation of GAN used to create or generate data with certain characteristics related to glucose in the body, particularly for medical applications) can be used to create synthetic glucose profiles, improving the accuracy of predictions in patients with type 1 diabetes. For personalised therapy, MOICVAE predicts drug sensitivity based on genomic data, while DRAGONET generates new candidates for the treatment of cancer and neurodegenerative diseases. In telemedicine services, ChatGPT and MedPaLM are used to automate consultations and generate reports, reducing patient service time. These results demonstrate the practical effectiveness of integrating AI into various stages of the medical process, from data collection to clinical decision support.

In practice, intelligent algorithms are effectively integrated into the functioning of virtual clinics and medical platforms. Virtual clinics Babylon Health and Ada Health use AI models for initial patient triage, automated symptom collection, and recommendations for further action (Cascella *et al.*, 2023). CardioAI and KardiaMobile systems use AI algorithms to analyse heart signals, detect arrhythmias, and monitor the cardiovascular system in real time (Thapa *et al.*, 2024). These examples demonstrate that virtual medicine has moved from experimentation to real-world clinical solutions, shaping accurate and safe digital medicine where AI becomes a "partner to the doctor" rather than a replacement.

Along with the development of clinical support systems, AI is becoming increasingly central in assisting patients before consulting a doctor. Virtual assistants and mobile applications with ML elements can be used to independently monitor basic physiological parameters such as body weight, blood pressure, heart rate, glucose levels, mood, and sleep quality. By analysing the dynamics of

these indicators, the system can detect early signs of abnormalities and generate personalised recommendations: to continue collecting additional data (for example, using smart sensors or home devices) or to consult a doctor of a specific profile if the identified trends are persistent. This approach creates conditions for preventive medicine, where AI not only supports doctors but also actively helps people stay healthy. In addition, Table 4 presents the key benefits of implementing AI technologies in a virtual healthcare system.

Table 4. Practical advantages of using AI in virtual medicine

Benefits	Description	Use
Improvement of diagnostic accuracy	Deep neural networks (CNN, Transformers) exceed the accuracy of doctors in image analysis, multimodal models combine visual, laboratory and text data (+18-25% to accuracy)	ADS-GAN – synthetic EHR data without loss of quality; MixEHR-G – 1,515 phenotypes from 1.3 million patients, HAIM – integration of different types of data, exceeding unimodal models
Optimisation of time and workload for doctors	AI automates routine tasks (image analysis, reports), reducing time and workload by 20-30%, while CDSS provides real-time guidance to doctors	cGAN – imputation of gaps in EHR, stable performance on MIMIC-III, ChatGPT – automation of dental consultations, reduction in assessment time
Expansion of access to quality healthcare	Telemedicine platforms, chatbots, and LLMs provide consultations in remote regions, language translation, and asynchronous patient support	cGAN – data generation for haematological research (≈ 7,000 samples), LLM – multilingual triage and communication with patients, CareCall bot – support for patients in remote locations
Reduction in the cost of medical care	Automated documentation, fewer readmissions and physical resources (10-15% savings)	Graph-GAN – synthetic EHRs with performance similar to supervised learning (10% labels), generative AI – reduction in data annotation costs

Note: ADS-GAN – adversarial domain-specific generative adversarial network; MixEHR-G – mixed-type electronic health records generative model; MIMIC-III – medical information mart for intensive care; CDSS – clinical decision support system
Source: compiled by the authors based on R. Shinde *et al.* (2022), L. Soenksen *et al*. (2022), M. Cascella *et al*. (2023), W. Abbaoui *et al*. (2024), I. Ghebrehiwet *et al.* (2024), H. Nilius *et al*. (2024), J. Pool *et al*. (2024), S. Thapa *et al.* (2024), E. Kumah (2025)

The practical benefits of AI in virtual medicine include increased accuracy, efficiency, accessibility, and cost-effectiveness of medical processes. In particular, deep neural networks (CNN, Transformers) exceed the average accuracy of radiologists, providing up to a 25% increase thanks to multimodal data integration (Soenksen *et al.*, 2022; Abbaoui *et al.*, 2024; Nilius *et al.*, 2024). cGAN and Graph-GAN models demonstrate effectiveness in generating synthetic medical records, which can be used to train models without disclosing personal data (Pool *et al.*, 2024; Kumah, 2025). ChatGPT and LLM platforms are used in telemedicine for automated consultations and linguistic adaptation, facilitating access to care in remote regions. The integration of AI tools into virtual medicine not only improves diagnostic accuracy but also contributes to the creation of a more efficient, inclusive, and economically sustainable healthcare system.

AI significantly improves the efficiency and accuracy of virtual medicine, optimises the workflow of medical staff, and expands access to quality services. At the same time, despite the advantages, the introduction of intelligent systems is accompanied by several challenges related to ethics, security, and legal regulation. AI systems make decisions that can directly affect human life and health, so issues of transparency, security, and accountability are relevant. Deep learning models function as "black boxes" and demonstrate high accuracy, but it is not always clear why a model makes a particular decision (van Kolfschooten & van Oirschot, 2024). This creates risks for medical practice, where every recommendation must be justified and reproducible.

The use of ML and DL models significantly improves diagnostic accuracy, analysis speed, and the effectiveness of early detection of pathologies. The integration of CNN, LSTM, and Transformers into virtual medicine systems can not only classify images but also generate predictive models of disease progression. This correlates with the conclusions of H. Sadr *et al.* (2025) that deep neural networks demonstrate a level of diagnostic accuracy comparable to or higher than that of expert radiologists, particularly in the analysis of medical images. The study emphasised the importance of hybrid architectures capable of combining different data modalities (EHR, images, time series) to achieve high generalisation ability. This confirmed the leading role of AI as a tool for improving the accuracy, speed and reliability of clinical diagnosis.

M. Khalifa & M. Albadawy (2024) demonstrated that DL algorithms, particularly CNN, U-Net, and ViT, significantly improve diagnostic accuracy by enabling automatic tissue segmentation and pathology detection with minimal human intervention. Multimodal models that combine medical imaging data, laboratory tests, and clinical records provide a more objective basis for decision-making and reduce the probability of errors in radiological practice. The results of this study support these conclusions: the introduction of DL algorithms into virtual medicine has improved the reliability of image analysis and reduced the time required to make a diagnosis. AI is a key factor in improving the accuracy, speed, and reliability of clinical diagnostics.

The results of the study showed that the use of intelligent algorithms in monitoring systems can be used for the detection of deviations in physiological indicators and the prediction of the development of complications in real time. The use of RNN, LSTM, and Transformer architectures facilitates accurate time series analysis and the formation of digital biomarkers for chronic diseases. This was consistent with the study by S. Shajari *et al.* (2023), demonstrating that wearable sensors integrated with AI provide continuous monitoring of patient status and improve the accuracy

of pathology detection. The combination of signals from wearable devices and intelligent algorithms creates a new model of digital health that prioritises prevention and early intervention, indicating the feasibility of using AI to create personalised dynamic health monitoring systems.

R.A. El Arab & O.A. Al Moosa (2025) have shown that the use of AI in medical practice can reduce overall service costs through automation, reduction of repeat hospitalisations, and resource optimisation. The study cited data on cost reductions in various AI application scenarios, from diagnostics to administrative services, demonstrating the positive budgetary impact of implementing intelligent systems. This was consistent with the presented study, which determined that the implementation of AI systems led to a significant reduction in costs, more efficient use of resources, and increased profitability of virtual medical services. AI not only improves clinical outcomes but also has a real economic impact that supports the sustainability of innovation in healthcare.

C.A. Gomez-Cabello *et al*. (2024) demonstrated that AI-based CDSS are being implemented in primary care, improving the quality of consultations and reducing the workload on doctors. The use of LLM, NLP methods, and ML algorithms automates the processes of triage, interpretation of examination results, and formulation of clinical recommendations, significantly reducing the cognitive load on doctors. In addition, the study noted that AI-CDSSs help improve interaction between patients and medical staff through adaptive interfaces and personalised information delivery, improving the quality of communication in telemedicine services. This approach was consistent with the results of the study: the integration of AI into telemedicine platforms and CDSS accelerated decision-making, reduced the burden on medical staff, and improved the accessibility of medical services. Thus, AI is transforming the virtual medicine system, ensuring higher diagnostic accuracy, faster clinical decisions, and cost-effective medical processes. The integration forms the basis of a human-centred, adaptive, and sustainable digital healthcare ecosystem of the new generation.

Regulatory and ethical aspects and prospects for the implementation of AI in virtual medicine

The issue of security in virtual medicine encompasses the protection of confidential medical data and the resilience of algorithms to external attacks. DL models can be vulnerable to adversarial attacks (deliberate changes in input data), leading to false diagnostic conclusions. In a medical context, such errors can have critical consequences, including misclassification of pathologies or incorrect clinical recommendations (Nilius *et al.*, 2024). Protecting patient privacy is the ethical foundation of digital medicine. Medical information belongs to the category of the most sensitive personal data, so its automated processing by artificial systems creates a potential threat of leakage, manipulation, or unauthorised access. Scenarios in which data is transferred via cloud services and integrated between several medical institutions or departmental information systems require critical attention (Abbaoui *et al.*, 2024).

Another significant ethical issue is the bias of training samples. When models are trained on unrepresentative data that predominantly covers patients of a certain gender, age or ethnic group, there is a risk of reproducing social inequalities. Such systems may demonstrate reduced accuracy or even generate discriminatory recommendations for other categories of patients, calling into question their fairness and reliability (Pool *et al.*, 2024). In addition, the principle of explainability is a prerequisite for the ethical use of AI models in clinical practice. The doctor must be able to justify the logic of the system's actions to the patient and explain the basis for the prognosis or recommendation. Despite significant progress in the development of Explainable AI methods, their reliability and reproducibility in real clinical settings remain limited.

The insufficient level of trust that doctors have in automated systems is also a significant barrier. Despite high accuracy rates in studies, clinicians tend to doubt the reliability of such solutions in real-life practice. The main reasons are the complexity of the models, the lack of explainability of the results, and the lack of standardised protocols for integrating AI into clinical processes. Trust in intelligent technologies is only formed when the system is transparent, validated, and provides a clear explanation of its decisions. At the same time, excessive reliance on algorithms is also dangerous: doctors should view AI as an assistant, not the final authority (Kumah, 2025). Therefore, modern 21st-century AI ethics emphasise the need to preserve the leading role of doctors in clinical decision-making, with automated systems performing a supporting function aimed at improving the quality and safety of medical care.

In the context of the growing role of patients as active participants in the healthcare process, it is advisable to consider the ethical and legal aspects of independent use of AI tools. Algorithms that provide advice or recommendations based on the user's personal data must be transparent in the logic of their conclusions and not create a false sense of "self-diagnosis". The collection of basic indicators (weight, blood pressure, mood, pulse, etc.) is only an auxiliary step that should guide a person to consciously seek professional help, rather than replace a doctor's consultation. Therefore, legal norms and standards for AI systems should cover not only clinical applications but also user scenarios for preliminary monitoring, ensuring a balance between convenience, safety, and responsibility.

One of the critical challenges of digital transformation in medicine remains the legal regulation of AI systems in clinical practice. Despite the rapid development of technology, regulatory mechanisms remain fragmented, especially in the area of safety and responsibility for medical decisions made with the participation of AI. At the EU level, a comprehensive regulatory framework has been developed that combines political and legal instruments. The European Commission (2024) document identifies five strategic areas for the development of digital medicine:

the safe integration of AI into clinical processes, including certification and risk assessment mechanisms; building trust in algorithms through transparency, explainability and controllability, creation of a European Health Data Space for the exchange of clinically relevant data sets between Member States, ethical and human-centred implementation of AI that guarantees the priority of human oversight, support for open data standards and interoperability in healthcare systems.

The document also defines ethical principles for the use of AI in medicine, such as patient safety as a key priority, explainability and transparency of algorithmic decisions, human oversight and final decision by a doctor, non-discrimination and prevention of algorithmic bias, and protection of personal data following the General Data Protection Regulation (2016). These principles are enshrined in the Artificial Intelligence Act (2024), which is the first EU legislation to systematically regulate AI. The AI Act classifies medical systems as high-risk and defines three groups of mandatory safety requirements: pre-deployment requirements (mandatory testing of the model for reliability and stability of results, verification of training data, assessment of potential harm to the patient), operational requirements (documentation of decision-making logic, traceability logs, continuous performance monitoring) and post-marketing surveillance (control of developers and operators after implementation, mandatory incident reporting, security system audits). In addition, the AI Act establishes the legal liability of developers and suppliers if it is proven that harm to the patient was caused by an algorithmic failure or non-compliance with data quality requirements. The European model combines ethical principles with specific technical and legal standards, ensuring transparency, traceability and human control at all stages of the AI system lifecycle.

In Ukraine, legal regulation is still in the early stages of development. The document "Roadmap for the Regulation of Artificial Intelligence in Ukraine" by the Ministry of Digital Transformation of Ukraine (2023) envisages the development of the Law of Ukraine "On Artificial Intelligence" and the harmonisation of approaches with the AI Act. The document defines the directions for development: introduction of risk classification for AI systems, development of safety and ethics assessment standards, creation of a national certification system for AI products, and establishment of the legal status of system developers and operators. At the same time, the Ukrainian strategy is conceptual in nature, as it does not contain definitions of the terms "medical algorithm", "clinical decision support system", or "automated recommendation". In contrast to the AI Act, it lacks specific requirements for testing, auditing, and liability in the healthcare sector. The Ukrainian approach is limited to emphasis on the future introduction of European standards, but without a practical mechanism for their implementation.

The European legal framework is characterised by a high level of specificity (requirements, classifications, security procedures), while the Ukrainian one is characterised by declarativeness and a lack of implementation tools. Both documents share a common value base of ethics, transparency and human-centredness, but differ in their level of detail and legal force. A comparison shows that the EU has already formed a multi-level regulatory model, where political strategy determines directions and principles, and the AI Act provides legal implementation. Ukraine is only approaching this system, maintaining its declared goals of harmonisation, but without legal mechanisms for security and accountability. To ensure the ethical implementation of AI in medicine, Ukraine needs to implement risk classification based on the AI Act model, introduce requirements for testing and auditing medical algorithms, legislate the principle of explainability and human control, and create a state body to oversee the safety of medical AI systems. The modern stage of AI evolution in virtual medicine is characterised by a transition from local, highly specialised solutions to complex, integrated and explainable systems that cover the entire clinical decision-making cycle, from data collection to the formation of diagnostic and therapeutic conclusions. Table 5 shows the promising areas of development for intelligent technologies.

Table 5. Conceptual directions for the formation of an intellectual healthcare system

Development area	Primary goal	Key technologies/examples	Challenges/trends
Explainable AI	Development of models capable of justifying decisions to increase the trust of doctors	Grad-CAM, LIME, SHAP for result interpretation; AI for phenotyping and in silico libraries	Balance between accuracy and interpretability; addressing the "black box"; ethical integration into clinical systems
Integration of LLM	Use of LLM for analysis, triage, and communication with patients	GPT-4/5, MedPaLM 2, BioMedLM; Retrieval-Augmented Generation (RAG); telemedicine, telepsychiatry	Control of bias; language localisation; responsible use and audit
Unification and standardisation of data	Ensuring interoperability and security of medical data	Federative learning; blockchain registries; multimodal integration (text, images, time series)	Unification of formats; personal data protection; elimination of "data silos"
Generative models and simulations	Creation of synthetic data and virtual scenarios for training and diagnostics	GAN, diffusion models, LLM-simulations; virtual patients and training	Data confidentiality; quality of synthetic samples; interpretability of models

Table 5. Continued

Development area	Primary goal	Key technologies/examples	Challenges/trends
Security and blockchain technologies	Protection of medical data and access control	Blockchain architectures for HER; AI audit and cyber resilience	Adversarial attacks; blockchain scalability; domain-specific decisions
Ethical and legal regulation	Provision of responsible use of AI in medicine	Regulation (EU) 2024/1689 (AI Act) Roadmap of AI regulation in Ukraine WHO, OECD, G7 standards	Certification of medical AI systems Harmonisation with the AI Act before 2027 Protection of patient rights and privacy

Note: Grad-CAM – Gradient-weighted Class Activation Mapping; LIME – Local Interpretable Model-agnostic Explanations; SHAP – Shapley Additive exPlanations

Source: compiled by the authors

The development of an intelligent healthcare system requires a combination of technological innovations with ethical and legal principles. The integration of Explainable AI changes the logic of diagnosis: doctors transition from "blind" use of the model to conscious interaction, forming the basis of responsible medical decision-making, where the algorithm enhances rather than replaces the specialist. LLMs expand the capabilities of telemedicine by providing personalised communication with patients and multilingual support, but require ethical control to prevent biased or erroneous conclusions.

Data standardisation through federated learning and blockchain paves the way for the creation of global clinical networks without compromising confidentiality, which will contribute to the formation of the European Health Data Space and its integration with Ukrainian systems. Generative models can be used to create digital twins of patients and virtual training, improving the quality of medical training without risk to patients. In the field of security and legal regulation, the key areas are the implementation of blockchain solutions for data protection and the harmonisation of legislation with the AI Act (2024/1689). Ukraine is gradually adapting the principles of certification, audit and transparency within the framework of the "Roadmap for AI Regulation". The further development of intelligent medicine requires a balanced combination of technological innovation, ethical responsibility and regulatory consistency. The identified conceptual directions, from explainable AI to the integration of blockchain and LLM, form the basis of a human-centred, safe and sustainable digital healthcare ecosystem.

The results of the study showed that the key barriers to the implementation of AI in medicine remain issues of ethical oversight, safety, and legal liability, which determine the level of trust in intelligent systems in virtual medicine. C. Mennella *et al.* (2024) analysed more than 150 studies and concluded that the main challenges for digital medicine are the lack of clear regulatory mechanisms, transparent model verification protocols and certification of medical AI systems. The study emphasised the need to create a regulatory framework that balances technological innovation with the ethical principles of patient autonomy, fairness, and privacy. These findings echo the conclusions of this study, which also emphasises the need to develop legal and ethical standards to minimise the risks of clinical AI applications. Both approaches confirm that without a reliable regulatory environment, the introduction of intelligent technologies into virtual medicine cannot be considered safe or ethically justified.

L. Tang *et al*. (2023) analysed the ethical aspects of the use of medical AI systems in a systematic review of empirical studies and determined that the main risks remain data bias, lack of model explainability, and uneven representation of socio-demographic groups in training samples. Such limitations lead to the reproduction of social inequalities in clinical recommendations and reduce user confidence in AI solutions. The results of this study are consistent with these findings, as they also found that the effectiveness of medical AI systems is determined not only by their level of accuracy, but also by their socio-ethical characteristics. Therefore, the elimination of algorithmic bias and the implementation of ethical standards are necessary conditions for the legitimate and safe use of AI in digital medicine.

In addition, Y. Ning *et al.* (2024) conducted a large-scale scoping review on the ethical challenges of using generative AI in medicine and developed an "Ethics checklist", a comprehensive system for assessing the risks of transparency, reliability, and accountability. The authors found that generative models, although they have significant potential for clinical decision support, can produce false or inaccurate conclusions, posing a danger to patients, particularly in the problem of determining responsibility for AI system errors between developers, medical institutions, and physician users. This correlates with the results of the current study on the risks of losing control over autonomous models and the need to regulate their clinical use. The opacity of generative AI can lead to ethical and legal conflicts in digital medicine; therefore, it is advisable to create audit, ethical monitoring and validation systems that will ensure the safe and accountable use of these technologies.

The results of the study indicated that explainability and user trust are key prerequisites for the successful implementation of AI in clinical practice. This was consistent with the study by K. Rasheed *et al.* (2022), stating that the development of Explainable AI is the only way to overcome the "black box" effect in deep neural networks. The study systematised model interpretation methods (LIME, SHAP, Grad-CAM) and showed how they help doctors verify the

system's predictions. Without explainability, AI solutions cannot be acceptable in a medical context where clinical responsibility depends on transparency. The development of Explainable AI is not only a technical but also an ethical imperative for digital medicine.

The study by A. Bathula *et al.* (2024) presented the concept of the "triangle of the future", based on the interaction of blockchain, AI, and digital medicine. The combination of these technologies provides decentralised storage of medical data, transparent auditing of user actions, and increased cyber resilience of medical systems. The study substantiated the role of blockchain as a trust tool that eliminates the risks of unauthorised access and data falsification, as well as enhances the security of AI models in telemedicine environments. This is consistent with current research that the integration of AI with blockchain is a strategic direction for the formation of secure and resilient virtual medical platforms. It is advisable to create hybrid frameworks that combine ML, NLP, and computer vision with blockchain verification mechanisms to protect data and prevent adversarial attacks. Thus, the key trend is the formation of a human-centred healthcare model, where AI does not replace but complements the clinical thinking of doctors, ensuring accuracy, transparency and personalisation of medical care. The synergy between technological, regulatory, and ethical components will determine the transition to a new generation of intelligent healthcare systems focused on the safe, open, and fair use of data in the global medical environment, particularly in Ukraine.

Conclusions

The results of the study demonstrated that AI methods are fundamental in the structural analysis of medical data, including EHR, laboratory indicators, and clinical records. SVM, Random Forest, and Decision Trees algorithms ensure the interpretability of decisions and work effectively with limited data volumes. Deep learning based on CNN, LSTM, and Transformer architectures has become the core of virtual medicine, enabling automatic detection of patterns in images, time series, and texts. The combination of ML and DL approaches forms adaptive models with high accuracy and clinical relevance. Generative models create simulations and "digital twins" of patients, while LLM, as part of telemedicine CDSS, accelerate triage, report preparation, and reduces the cognitive load on physicians. The integration of different modalities through cross-modal attention mechanisms provides a comprehensive clinical picture, reducing the risk of wrong decisions in complex cases (comorbidity, rare conditions). The economic impact of AI is reflected in a reduction in readmissions and data annotation/processing costs (through synthetic data and automation), which increases the profitability of digital services.

Key risks remain the vulnerability of models to adversarial attacks, bias in training samples, and the opacity of deep networks, which undermines trust in clinical decisions and necessitates increased explainability and independent auditing. To scale virtual medicine across clinical institutions, the unification of clinical and laboratory data, the interoperability of electronic records, and federated training of AI models are key. Combining this with blockchain ensures data immutability and transparency of access to medical informatiofdn. The EU regulatory vector sets requirements for transparency, certification, and risk management for medical AI systems. Ukraine should harmonise its legislation with the AI Act through the phased implementation of assessment, audit, and ethical control procedures. The priority steps for implementing AI in virtual medicine are Explainable AI approaches in critical tasks, RAG architectures for LLM with controlled knowledge sources, data policies (federated learning + blockchain) and continuous ethical monitoring. Further research should address the integration of explanatory and generative models, the development of secure and transparent AI frameworks, and the creation of a global regulatory and ethical ecosystem for virtual medicine.

Acknowledgements

None.

Funding

The study was not funded.

Conflict of Interest

None.

References

[1] Abbaoui, W., Retal, S., El Bhiri, B., Kharmoum, N., & Ziti, S. (2024). Towards revolutionizing precision healthcare: A systematic literature review of artificial intelligence methods in precision medicine. *Informatics in Medicine Unlocked*, 46, article number 101475. doi: 10.1016/j.imu.2024.101475.

[2] Abbas, S.R., Seol, H., Abbas, Z., & Lee, S.W. (2025). Exploring the role of artificial intelligence in smart healthcare: A capability and function-oriented review. *Healthcare*, 13(14), article number 1642. doi: 10.3390/healthcare13141642.

[3] Alkhanbouli, R., Almadhaani, H.M., Alhosani, F., & Simsekler, M.C. (2025). The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions. *BMC Medical Informatics and Decision Making*, 25, article number 110. doi: 10.1186/s12911-025-02944-6.

[4] Artificial Intelligence Act. (2024, August). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689.

[5] Bathula, A., *et al*. (2024). Blockchain, artificial intelligence, and healthcare: The tripod of future – a narrative review. *Artificial Intelligence Review*, 57, article number 238. doi: 10.1007/s10462-024-10873-5.

[6] Boychenko, O., & Bubliy, T. (2024). Prospects for the use of artificial intelligence in the medical field. *Current Issues in Modern Medicine: Bulletin of the Ukrainian Medical Stomatological Academy*, 24(3), 137-139. doi: 10.31718/2077-1096.24.3.137.

[7] Cascella, M., Scarpati, G., Bignami, E.G., Cuomo, A., Vittori, A., Di Gennaro, P., Crispo, A., & Coluccia, S. (2023). Utilizing an artificial intelligence framework (conditional generative adversarial network) to enhance telemedicine strategies for cancer pain management. *Journal of Anesthesia, Analgesia and Critical Care*, 3, article number 19. doi: 10.1186/s44158-023-00104-8.

[8] Clark, H.B., Egger, J., & Duffy, V.G. (2024). AI in healthcare and medicine: A systematic literature review and reappraisal. In V.G. Duffy (Ed.), *Digital human modeling and applications in health, safety, ergonomics and risk management. HCII 2024. Lecture notes in computer science* (Vol. 14710, pp. 251-270). Cham: Springer. doi: 10.1007/978-3-031-61063-9_17.

[9] El Arab, R.A., & Al Moosa, O.A. (2025). Systematic review of cost effectiveness and budget impact of artificial intelligence in healthcare. *npj Digital Medicine*, 8, article number 548. doi: 10.1038/s41746-025-01722-y.

[10] European Commission. (2024). *Artificial intelligence in healthcare*. Retrieved from https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en.

[11] General Data Protection Regulation. (2016, April). Retrieved from https://gdpr-text.com.

[12] Ghebrehiwet, I., Zaki, N., Damseh, R., & Mohamad, M.S. (2024). Revolutionizing personalized medicine with generative AI: A systematic review. *Artificial Intelligence Review*, 57, article number 128. doi: 10.1007/s10462-024-10768-5.

[13] Gomez-Cabello, C.A., Borna, S., Pressman, S., Haider, S.A., Haider, C.R., & Forte, A.J. (2024). Artificial-intelligence-based clinical decision support systems in primary care: A scoping review of current clinical implementations. *European Journal of Investigative Health Psychology and Education*, 14(3), 685-698. doi: 10.3390/ejihpe14030045.

[14] Guo, X., & Li, Y. (2024). Intelligent health in the IS area: A literature review and research agenda. *Fundamental Research*, 4(4), 961-971. doi: 10.1016/j.fmre.2023.04.008.

[15] Khalifa, M., & Albadawy, M. (2024). AI in diagnostic imaging: Revolutionising accuracy and efficiency. *Computer Methods and Programs in Biomedicine Update*, 5, article number 100146. doi: 10.1016/j.cmpbup.2024.100146.

[16] Khan, M.M., Shah, N., Shaikh, N., Thabet, A., Alrabayah, T., & Belkhair, S. (2025). Towards secure and trusted AI in healthcare: A systematic review of emerging innovations and ethical challenges. *International Journal of Medical Informatics*, 195, article number 105780. doi: 10.1016/j.ijmedinf.2024.105780.

[17] Korotka, V.O., & Mokrynskyi, V.A. (2024). Technologies of artificial intelligence in modern medicine: Implementation and issues. *Digital Medicine*, 163(5), 119-121. doi: 10.32471/umj.1680-3051.163.257497.

[18] Kumah, E. (2025). Artificial intelligence in healthcare and its implications for patient-centered care. *Discover Public Health*, 22, article number 524. doi: 10.1186/s12982-025-00924-9.

[19] Mennella, C., Maniscalco, U., De Pietro, G., & Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon*, 10(4), article number e26297. doi: 10.1016/j.heliyon.2024.e26297.

[20] Ministry of Digital Transformation of Ukraine. (2023). *Regulation of artificial intelligence in Ukraine: Presenting a roadmap*. Retrieved from https://thedigital.gov.ua/news/technologies/regulyuvannya-shtuchnogo-intelektu-v-ukraini-prezentuemo-dorozhnyu-kartu.

[21] Mizna, S., Arora, S., Saluja, P., Das, G., & Alanesi, W.A. (2025). An analytic research and review of the literature on the practice of artificial intelligence in healthcare. *European Journal of Medical Research*, 30, article number 382. doi: 10.1186/s40001-025-02603-6.

[22] Nilius, H., Tsouka, S., Nagler, M., & Masoodi, M. (2024). Machine learning applications in precision medicine: Overcoming challenges and unlocking potential. *TrAC Trends in Analytical Chemistry*, 179, article number 117872. doi: 10.1016/j.trac.2024.117872.

[23] Ning, Y., *et al*. (2024). Generative artificial intelligence and ethical considerations in health care: A scoping review and ethics checklist. *The Lancet Digital Health*, 6(11), E848-E856. doi: 10.1016/S2589-7500(24)00143-2.

[24] Pool, J., Indulska, M., & Sadiq, S. (2024). Large language models and generative AI in telehealth: A responsible use lens. *Journal of the American Medical Informatics Association*, 31(9), 2125-2136. doi: 10.1093/jamia/ocae035.

[25] Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2022). Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine*, 149, article number 106043. doi: 10.1016/j.compbiomed.2022.106043.

[26] Sadr, H., *et al*. (2025). Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: A comprehensive review of machine learning and deep learning approaches. *European Journal of Medical Research*, 30, article number 418. doi: 10.1186/s40001-025-02680-7.

[27] Shajari, S., Kuruvinashetti, K., Komeili, A., & Sundararaj, U. (2023). The emergence of AI-based wearable sensors for digital health technology: A review. *Sensors*, 23(23), article number 9498. doi: 10.3390/s23239498.

[28] Shinde, R., Patil, S., Kotecha, K., Potdar, V., Selvachandran, G., & Abraham, A. (2022). Securing AI-based healthcare systems using blockchain technology: A state-of-the-art systematic literature review and future research directions. *ArXiv*. doi: 10.48550/arXiv.2206.04793.

[29] Soenksen, L.R., Ma, Y., Zeng, C., Boussioux, L.D., Villalobos Carballo, K., Na, L., Wiberg, H.M., Li, M.L., Fuentes, I., & Bertsimas, D. (2022). Integrated multimodal artificial intelligence framework for healthcare applications. *ArXiv*. doi: 10.48550/arXiv.2202.12998.

[30] Sofilkanych, N., Vesova, O., Kaminskyy, V., & Kryvosheieva, A. (2023). The impact of artificial intelligence on Ukrainian medicine: Benefits and challenges for the future. *Futurity Medicine*, 2(4), 28-39. doi: 10.57125/FEM.2023.12.30.04.

[31] Tang, L., Li, J., & Fantus, S. (2023). Medical artificial intelligence ethics: A systematic review of empirical studies. *Digital Health*, 9, 1-22. doi: 10.1177/20552076231186064.

[32] Thapa, S., Fakiraswamimath, A.P., Zuluaga, M., Kumar, A.R., Ramesh, K., & Yadav, S. (2024). The role of artificial intelligence in personalized medicine: Current trends and future directions. *Frontiers in Health Informatics*, 13(3), 3830-3841.

[33] van Kolfschooten, H., & van Oirschot, J. (2024). The EU Artificial Intelligence Act (2024): Implications for healthcare. *Health Policy*, 149, article number 105152. doi: 10.1016/j.healthpol.2024.105152.

Використання інтелектуальних алгоритмів у комп'ютерних системах віртуальної охорони здоров'я: від діагностики до персоналізованого лікування

Микола Хрульов

Аспірант
Черкаський державний технологічний університет
18006, б-р Шевченка, 460, м. Черкаси, Україна
https://orcid.org/0000-0001-8532-0967

Тетяна Миронюк

Кандидат технічних наук, доцент
Черкаський державний технологічний університет
18006, б-р Шевченка, 460, м. Черкаси, Україна
https://orcid.org/0000-0002-7588-1055

Анотація. Метою дослідження було теоретично обґрунтувати підходи до ефективного впровадження інтелектуальних алгоритмів у віртуальну медицину. Методологія ґрунтувалась на теоретичному, аналітичному та нормативно-прогнозному аналізі ефективності й розвитку інтелектуальних технологій у цифровій охороні здоров'я. Встановлено, що штучний інтелект (ШІ) трансформує підходи до збору, аналізу та використання медичних даних. Віртуальна медицина застосовує машинне й глибоке навчання для діагностики, прогнозування та персоналізованого лікування, підвищуючи точність рішень і зменшуючи навантаження на лікарів. Методи машинного навчання ефективні для роботи з електронними медичними записами та лабораторними даними, тоді як глибоке навчання формує основу віртуальної медицини, автоматизуючи аналіз великих обсягів інформації. Генеративні моделі створюють синтетичні медичні дані й клінічні сценарії, підтримуючи розвиток персоналізованої медицини та концепції «цифрових двійників». Мультимодальні системи поєднують різні типи даних, забезпечуючи комплексний аналіз стану пацієнта й точніші клінічні прогнози. Переваги впровадження ШІ у підвищенні точності діагностики на 18–25 %, зменшенні часу роботи лікарів на 20–30 %, розширенні доступу до медицини у віддалених регіонах, зниження вартості медичних послуг. Основними ризиками є проблеми безпеки даних, пояснюваності, етики, упередженості та довіри лікарів, що зумовлює потребу у прозорості, контролі й правовому регулюванні. У Європейському Союзі діє спеціальне законодавство, яке встановлює вимоги до безпеки та прозорості медичних ШІ-систем, тоді як в Україні нормативна база перебуває на етапі формування. Для вдосконалення віртуальної медицини доцільно впровадити пояснюваний ШІ, інтегрувати Large Language Models із захистом даних, застосовувати федеративне навчання, генеративні симуляції та блокчейн із дотриманням етичних і правових стандартів. Результати дослідження можуть бути використані фахівцями при прийнятті рішень щодо вибору і застосування інтелектуальних алгоритмів у медичних закладах, дослідницьких центрах т а ІТ-сфері охорони здоров'я

Ключові слова: мультимодальна аналітика медичних даних; цифровий моніторинг; генеративні моделі для симуляцій; пояснюваність клінічних рішень; дистанційна медицина; безпека та приватність медичних даних

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ІНЖЕНЕРІЯ

Науково-технічний журнал

Том 22, № 3, 2025

Заснований у 2004 р. Виходить 3 рази на рік

Оригінал-макет видання виготовлено
у редакційно-видавничому відділі Вінницького національного технічного університету.

Відповідальний редактор:
В. Белзецька

Підписано до друку 23.12.2025 р. Формат 60*84/8
Умовн. друк. арк. 22,8
Наклад 100 примірників

Адреса видавництва:
Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна
Тел: +38 (0432) 560848
Факс: +38 (0432) 465772
E-mail: info@itce.vn.ua
https://itce.vn.ua/uk

INFORMATION TECHNOLOGIES AND COMPUTER ENGINEERING

Scientific and Technical Journal

Vol. 22, No. 3, 2025

Founded in 2004. Published three times per year

The original layout of the publication is made
in the publishing department of Vinnytsia National Technical University

Managing editor:
V. Belzetska

Signed for print 23.12.2025. Format 60*84/8
Conventional printed pages 22.8
Circulation 100 copies

Publishing Address:
Vinnytsia National Technical University
21021, 95 Khmelnytske Shose Str., Vinnytsia, Ukraine
Telephone: +38 (0432) 560848
Fax: +38 (0432) 465772
E-mail: info@itce.vn.ua
https://itce.vn.ua/en