

VOLUME 1

|

NUMBER 1

|

JUNE 2026

JOURNAL OF

SECURE DIGITAL TECHNOLOGIES

A PEER-REVIEWED JOURNAL

BRIDGING TECHNOLOGY, LAW AND TRUST

IN THE DIGITAL AGE

DGSC, Astana, Kazakhstan

Journal of Secure Digital Technologies

Volume 1 (№1) June 2026

Sertificate № KZ19VPY00128988, date of issue: 25.12.2025

*The journal is published 4 times a year
(March, June, September, December)*

EDITORIAL TEAM

Editor-in-Chief

Dr. Leila Rzayeva, Association prof., PhD, Digital government support center (Kazakhstan)

Deputy Editor-in-Chief

Dr. Ali Myrzatay, PhD, Digital government support center (Kazakhstan)

Executive Secretary

Dr. Abdenova Zhuldyz, PhD, Digital government support center (Kazakhstan)

EDITORIAL BOARD MEMBERS (KAZAKHSTAN)

Seitkulov E.N. – Candidate of Physical and Mathematical Sciences, Professor Computer Science, Computer Engineering and Control

Abdenova G.A. – PhD in Automation and Control, Candidate of Technical Sciences, Professor L.N. Gumilyov Eurasian National University

Atanov S.K. – Doctor of Technical Sciences, Professor.

Sergazin G.K. – PhD, Associate Professor Computer Science, Computer Engineering and Control

Satybaldina D.Zh. – Candidate of Physical and Mathematical Sciences, Professor Director, Research Institute of Information Security and Cryptology, L.N. Gumilyov Eurasian National University

INTERNATIONAL EDITORIAL BOARD MEMBERS

Ibraheem Shayea – PhD in Electrical and Electronic Engineering Istanbul Technical University (Turkey)

Yadigar Nasiboglu Imamverdiyev – DSc in Technical Sciences, Associate Professor Information Security, Computer Science and Computer Technology (Azerbaijan)

Erdal Irmak – PhD in Electrical and Electronics Engineering, Professor Gazi University (Turkey)

Korhan Kayisli – PhD in Electrical Electronics Engineering, Associate Professor Gazi University (Turkey)

Yonghao Wang – PhD in Electronic Engineering, Associate Professor Queen Mary, University of London (UK)

Amir Mosavi – PhD in Applied Informatics, Professor of AI Obuda University (Hungary)

Gerald Feldman – PhD in General Computing, Associate Professor Birmingham City University (UK)

Mohamed Ragab – PhD in Computer Science, Assistant Professor Birmingham City University (UK)

Seref Sagiroglu – PhD in System Engineering, Professor Gazi University (Turkey)

MANAGING EDITOR

Dr. Ali Myrzatay, PhD, Digital government support center (Kazakhstan)

TECHNICAL EDITOR

Dr. Ali Myrzatay, PhD, Digital government support center (Kazakhstan)

CONTENTS

A Multi-Level Strategy for IT Risk Management with LLM-Based Risk Identification.....	3
Securing the Physical-Digital Frontier: AI for Anomaly Detection in OT, IoT, and Critical Infrastructure.....	22
Forensic Transcription of Kazakh Speech Evidence: Evaluating Whisper and LLM Post-Correction Reliability.....	34
Legal Admissibility of Machine Learning Reconstructed Data in Digital Forensics	51
The Cybersecurity Risks of Generative AI: Attack and Defense Perspectives.....	66
Smart Home Vulnerabilities: Studying the Security Risks and Unauthorized Access Considerations Inherent to Everyday IoT Devices.....	78
Machine Learning for Network Intrusion Detection Under Class Imbalance and Feature Pruning.....	92

Fakhreddin F. Khassanov

M.Sc. student, Department of Information
Technology, Faculty of Engineering and Information
Technology
khassanov.research@example.com (placeholder) /
ORCID: 0000-0000-0000-0000 (placeholder)
Kazakh University of Technology and Business
named after K. Kulazhanov, Kazakhstan

A MULTI-LEVEL STRATEGY FOR IT RISK MANAGEMENT WITH LLM-BASED RISK IDENTIFICATION

Abstract: Digital transformation has expanded the range of risks facing IT projects worldwide, yet fragmented, non-integrated adoption of governance and risk standards such as ISO 31000, PMBOK 7, COBIT 2019 and NIST SP 800-30 leaves a structural gap between corporate governance and day-to-day project risk control, a gap that is especially acute for small and medium teams in emerging markets such as Kazakhstan. This study develops and empirically validates a Multi-Level Integrated Risk Architecture (MIAR) that embeds strategic, tactical and operational risk controls into the IT project lifecycle; introduces a normalized Risk Exposure Reduction Index (RERI) with criticality weighting and cost/timing correction factors estimated with bootstrap confidence intervals; and implements IT Risk Analyzer, a multi-tenant web service that uses a large language model (Claude API) with chain-of-thought and few-shot prompting to automatically extract structured risk registers from technical-specification text. The AI module was evaluated against expert-built ground truth, cross-verified with the NIST SP 800-30, OWASP Top 10 and PMBOK 7 catalogues, on five real technical specifications, and the complete strategy was validated through a nine-month quasi-experiment on a government digital-forensics IT project (budget $\approx$ 186 million KZT). The AI module reached precision 78.2%, recall 86.4% and F1 81.9%. Applying MIAR to the case project reduced weighted risk exposure by $RERI_{base} = 0.641$ (95% CI [0.563, 0.719]) and $RERI_{ext} = 0.743$, reached a Risk Management Index of 90%, and produced an estimated methodology ROI of 601%, corresponding to approximately 43.5 million KZT in avoided expected losses. These results indicate that combining an integrated multi-standard governance architecture with LLM-assisted risk identification yields measurable, reproducible reductions in risk exposure and provides a deployable instrument for resource-constrained IT teams and public-sector projects.

Keywords: *IT risk management; risk governance integration; large language models; risk exposure reduction index; Monte Carlo simulation; prompt engineering; government IT projects; digital forensics compliance*

Многоуровневая стратегия управления ИТ-рисками с идентификацией рисков на основе больших языковых моделей (LLM)

Аннотация: Цифровая трансформация расширила спектр рисков, с которыми сталкиваются ИТ-проекты по всему миру. При этом фрагментарное, неинтегрированное внедрение стандартов управления и рисков, таких как ISO 31000, PMBOK 7, COBIT 2019 и NIST SP 800-30, оставляет структурный разрыв между корпоративным управлением и повседневным контролем проектных рисков. Этот разрыв особенно остро ощущается в малых и средних командах на развивающихся рынках, таких как Казахстан. В данном исследовании разрабатывается и эмпирически обосновывается Многоуровневая интегрированная архитектура рисков (MIAR), которая встраивает стратегический, тактический и операционный контроль рисков в жизненный цикл ИТ-проекта. В работе вводится нормированный Индекс снижения подверженности рискам (RERI) с весовыми коэффициентами критичности и корректирующими факторами стоимости/сроков, оцененными с помощью бутстреп-доверительных интервалов. Также реализован IT Risk Analyzer — мультиарендный веб-сервис, использующий большую языковую модель (Claude API) с цепочкой рассуждений (chain-of-thought) и обучением на нескольких примерах (few-shot prompting) для автоматического извлечения структурированных реестров рисков из текстов технических спецификаций. Работа ИИ-модуля оценивалась в сравнении с эталонными данными, составленными экспертами, и сопоставлялась с каталогами NIST SP 800-30, OWASP Top 10 и PMBOK 7 на примере пяти реальных технических спецификаций. Вся стратегия целиком была валидирована в ходе девятимесячного квазиэксперимента в государственном ИТ-проекте по цифровой криминалистике (бюджет ≈ 186 млн тенге). ИИ-модуль достиг точности (precision) 78,2%, полноты (recall) 86,4% и F1-меры 81,9%. Применение MIAR к пилотному проекту снизило взвешенную подверженность рискам на $RERI_{base} = 0,641$ (95% ДИ) и $RERI_{ext} = 0,743$, позволило достичь Индекса управления рисками в 90% и обеспечило расчетную окупаемость методологии (ROI) на уровне 601%, что соответствует примерно 43,5 млн тенге предотвращенных ожидаемых потерь. Эти результаты показывают, что сочетание интегрированной многостандартной архитектуры управления с идентификацией рисков при помощи LLM обеспечивает измеримое и воспроизводимое снижение подверженности рискам.

Ключевые слова: *управление ИТ-рисками; интеграция управления рисками; большие языковые модели; индекс снижения подверженности рискам; симуляция Монте-Карло; инженерия промтов; государственные ИТ-проекты; соответствие требованиям в цифровой криминалистике*

Үлкен Тілдік Модельдерге (LLM) Негізделген Тәуекелдерді Анықтау Арқылы Ат-Тәуекелдерді Басқарудың Көпдеңгейлі Стратегиясы

Аңдатпа: Цифрлық трансформация бүкіл әлемдегі АТ-жобалар алдында тұрған тәуекелдер ауқымын кеңейтті. Дегенмен, ISO 31000, PMBOK 7, COBIT 2019 және NIST SP 800-30 сияқты басқару және тәуекелдер стандарттарын үзінді-үзінді, біріктірілмей енгізу корпоративтік басқару мен күнделікті жобалық тәуекелдерді бақылау арасында құрылымдық алшақтықты қалдырады. Бұл алшақтық әсіресе Қазақстан сияқты дамып келе жатқан нарықтардағы шағын және орта командалар үшін өте өзекті. Бұл зерттеуде АТ-жобасының өмірлік цикліне стратегиялық, тактикалық және операциялық тәуекелдерді бақылауды ендіретін Көпдеңгейлі интеграцияланған тәуекелдер архитектурасы (MIAR) әзірленген және эмпирикалық түрде дәлелденген. Жұмыста бутстреп сенімділік аралықтарымен бағаланған маңыздылық салмақтары және шығын/уақыт түзету коэффициенттері бар нормаланған Тәуекел деңгейін төмендету индексі (RERI) енгізіледі. Сондай-ақ техникалық ерекшеліктер (спецификациялар) мәтінінен құрылымдалған тәуекелдер тізілімін автоматты түрде шығарып алу үшін пайымдаулар тізбегі (chain-of-thought) мен бірнеше мысалдар арқылы оқытуды (few-shot prompting) қолданатын, үлкен тілдік модельге (Claude API) негізделген IT Risk Analyzer көппайдаланушылық веб-сервисі іске асырылған. Жасанды интеллект (ЖИ) модулі бес шынайы техникалық ерекшелік негізінде сарапшылар жасаған эталондық мәліметтермен бағаланып, NIST SP 800-30, OWASP Top 10 және PMBOK 7 каталогтарымен салыстырыла тексерілді. Ал толық стратегия мемлекеттік цифрлық криминалистикалық АТ-жобасындағы (бюджеті ≈186 млн теңге) тоғыз айлық квазиэксперимент арқылы мақұлданды. ЖИ модулі 78,2% дәлдікке (precision), 86,4% толықтыққа (recall) және 81,9% F1-өлшеміне қол жеткізді. Пилоттық жобаға MIAR-ды қолдану өлшенген тәуекел деңгейін $RERI_{base} = 0,641$ (95% СИ) және $RERI_{ext} = 0,743$ -ке азайтты, Тәуекелдерді басқару индексін 90%-ға жеткізді және әдіснаманың болжамды қайтарымдылығын (ROI) 601% деңгейінде қамтамасыз етті, бұл шамамен 43,5 млн теңге алдын алынған күтілетін шығындарға сәйкес келеді. Бұл нәтижелер интеграцияланған көпстандартты басқару архитектурасын LLM көмегімен тәуекелдерді анықтаумен біріктіру тәуекел деңгейін өлшенетін және қайталанатын түрде төмендетуге мүмкіндік беретінін көрсетеді.

Түйін сөздер: *АТ-тәуекелдерді басқару; тәуекелдерді басқаруды интеграциялау; үлкен тілдік модельдер; тәуекел деңгейін төмендету индексі; Монте-Карло симуляциясы; промт инженериясы; мемлекеттік АТ-жобалар; цифрлық криминалистикадағы талаптарға сәйкестік*

1. Introduction

Over the past decade, digital transformation has expanded the range and complexity of risks faced by organizations delivering IT projects. According to the Standish Group Chaos Report, only 31% of IT projects are completed on time and within budget worldwide; about 52% overrun schedule or budget, and 17% are abandoned outright, with global losses

from failed IT projects exceeding USD 2.5 trillion annually [1]. Kazakhstan is undergoing an analogous transition under the state “Digital Kazakhstan” programme, with e-government, fintech, EdTech and MedTech expanding rapidly; the national IT market exceeded 1.2 trillion KZT in 2024 [2]. In this context, IT project risk becomes not only a commercial concern but a social one, particularly for government-funded initiatives.

A persistent paradox characterizes current practice: despite mature international standards (ISO 31000 [3], PMBOK 7 [4], NIST SP 800-30 [5], COBIT 2019 [6] and COSO ERM [7]), most delivery teams apply them partially or not at all. These standards were developed independently, each with its own scope (enterprise governance, project management, security risk assessment), and their integration into a single operating system of risk control remains an unresolved practical problem, a difficulty long noted in the broader risk-management literature [8]. Foundational software engineering literature has long recognized that schedule and quality risk in software projects is structural rather than incidental [9–11], yet translating this insight into an integrated, measurable governance-to-operations pipeline has received comparatively little attention.

A further source of difficulty is risk identification itself. The conventional process depends on experienced experts, is time-consuming, and remains subjective. The emergence of large language models (LLMs) opens the possibility of automatically identifying risks from the text of a technical specification (ToR), which is particularly useful for small teams and startups without a dedicated risk manager. Recent work has begun to explore LLM- and NLP-based risk identification with encouraging but uneven results: rule-based and classical NLP approaches remain limited in recall and adaptability, while transformer-based approaches show higher accuracy at the cost of dependence on external APIs and non-deterministic output [12–15]. Prompting techniques such as chain-of-thought reasoning and few-shot calibration have been shown to substantially improve the reliability of LLM output on structured reasoning tasks without requiring a labeled training corpus [16–17], which is attractive when no domain-specific risk dataset exists, as is presently the case in Kazakhstan.

In the Kazakhstani academic literature specifically, rule-based and matrix-based approaches to IT and information-security risk assessment have been proposed [18–20], but none combine an integrated cross-standard governance architecture with a normalized, comparable effectiveness metric and an LLM-based identification tool validated on a real project. A preliminary, narrower account of the governance-integration component of the present strategy appeared in an earlier publication by the author [21]; the present paper substantially extends that work with the RERI/RMI measurement framework, the AI identification module, and full quasi-experimental case validation. This combination of a mature international evidence base on one side, and a chronic practical lag (fragmented tooling, unmeasured risk-management outcomes, and an underdeveloped local body of work) on the other, constitutes the research gap this study addresses.

This study addresses the following research questions: RQ1: can a single integrated architecture be designed that removes the structural gaps between ISO 31000, PMBOK 7,

COBIT 2019 and NIST SP 800-30 and is applicable in real IT projects? RQ2: can current large language models identify IT risks from technical-specification text with accuracy comparable to expert judgment ($F1 \geq 75\%$)? RQ3: does the resulting strategy produce measurable, verifiable risk-exposure reduction under real project conditions ($RERI \geq 0.5$, $RMI \geq 80\%$)?

The scientific novelty of this work consists in three contributions that, to the authors' knowledge, have not previously been combined in a single validated framework: (1) the Multi-Level Integrated Risk Architecture (MIAR), a three-tier model that closes both a vertical gap (strategy–project–system) and a horizontal gap (risk management as a parallel activity rather than one embedded in the lifecycle) among COBIT 2019, ISO 31000, PMBOK 7 and NIST SP 800-30; (2) the Risk Exposure Reduction Index (RERI), a normalized, criticality-weighted metric with cost and timing correction factors and bootstrap confidence intervals, which distinguishes it from existing effectiveness measures such as Residual Risk or Risk Burn-down; and (3) a verified prompting framework for Claude-API-based automatic risk identification that achieves accuracy comparable to published international benchmarks [14, 15] without requiring a labeled training corpus.

Contributions. The concrete contributions of this paper are:

- MIAR, a three-level (strategic–tactical–operational) integrated risk-governance architecture, together with a seven-category, 24-subcategory IT risk taxonomy;
- RERI, a normalized risk exposure reduction index (base and extended forms) with a documented bootstrap confidence-interval procedure, together with the complementary Risk Management Index (RMI);
- IT Risk Analyzer, a production multi-tenant web service that couples an LLM-based risk-identification module (precision 78.2%, recall 86.4%, $F1$ 81.9%, 18–25 s generation time) with a Monte Carlo simulation engine, a Kanban workflow, and Jira/Azure DevOps/Slack integrations;
- empirical validation on a real government IT project, showing $RERI_{base} = 0.641$, $RERI_{ext} = 0.743$, $RMI = 90\%$, an estimated methodology ROI of 601%, and approximately 43.5 million KZT in avoided expected losses.

The remainder of the paper is organized as follows: Section 2 presents the research design, the MIAR architecture, the RERI/RMI methodology and the architecture of the IT Risk Analyzer system; Section 3 reports the quantitative results of the AI-module evaluation and of the case-project validation; Section 4 discusses the findings against the research questions and prior work; Section 5 concludes the paper.

2. Methods and Materials

This study combines design science research (DSR) for the construction of the MIAR architecture and the IT Risk Analyzer artifact with a single-case quasi-experimental design for empirical validation. DSR was chosen over a purely positivist experimental approach because the object of study is, in part, an artifact to be built and evaluated, not only a

phenomenon to be observed [22]. The following subsections describe (2.1) the overall research design and case context, (2.2) the MIAR architecture and risk taxonomy, (2.3) the risk assessment methodology, (2.4) the RERI and RMI metrics, (2.5) the architecture of the IT Risk Analyzer system, and (2.6) the evaluation design for the AI module and the case study.

2.1 Research Design and Case Study Context

The overall design is mixed: a systematic comparative analysis of five governance/risk standards (ISO 31000, PMBOK 7, NIST SP 800-30, COBIT 2019, COSO ERM) was used to identify structural gaps and inform the MIAR architecture; prompt engineering (chain-of-thought and few-shot calibration) was used to adapt an LLM to risk identification without a labeled corpus, which is appropriate given the absence of a public Kazakhstani IT-risk dataset; and a quasi-experimental pre/post comparison within a single real project was used to validate the complete strategy, since no parallel control project without the methodology was available.

The case project, referred to here as “NIC CyberTech” per the confidentiality requirements of the client organization, is a government-funded digital-forensics software project (budget $\approx$ 186 million KZT, 14-month duration). The author acted as systems analyst / project manager throughout. Risk exposure was recorded at two checkpoints: a baseline profile at the start of active risk management, and a residual profile after mitigation measures were implemented, over a nine-month observation window (ten months of active development in total). The registry comprised 30 identified risks (3 Critical, 5 High, 17 Medium, 5 Low).

Threats to validity were addressed as follows. For internal validity, probability/impact estimates were recorded before the AI module's output was reviewed (a pre-registered scoring criterion), and residual estimates were agreed jointly with the client rather than set unilaterally by the author. For external validity, the case project is a government project with heightened forensic-evidentiary requirements, so the specific risk composition and RERI value should not be assumed to generalize directly to commercial projects; the taxonomy and index themselves are designed to be transferable. For construct validity, RERI is always interpreted jointly with the absolute residual score of critical risks, since a high RERI alone does not guarantee an acceptable absolute risk level.

2.2 Multi-Level Integrated Risk Architecture (MIAR)

The comparative analysis of ISO 31000, PMBOK 7, NIST SP 800-30, COBIT 2019 and COSO ERM identified two systematic gaps. A vertical gap separates enterprise-level governance (risk appetite, board oversight; COBIT 2019, COSO ERM) from project-level risk registers (ISO 31000, PMBOK 7) and from system-level operational controls (NIST SP 800-30, OWASP); each standard is typically applied in isolation by a different organizational function. A horizontal gap treats risk management as an activity running parallel to the project lifecycle rather than embedded within it, so that risk reviews occur periodically rather than at each lifecycle transition.

MIAR closes both gaps with a three-level architecture aligned with the project lifecycle axis (initiation, planning, development, testing, deployment, operation). Level 1 (Strategic) covers risk appetite, policy and board-level oversight, grounded in COBIT 2019 and COSO ERM. Level 2 (Tactical) covers the project risk register and response planning, grounded in ISO 31000 and PMBOK 7, and is owned by the project management office. Level 3 (Operational) covers technical controls such as secure coding, SAST/DAST and monitoring, grounded in NIST SP 800-30 [5] and OWASP [23], and is owned by delivery teams. Each level feeds the level above it, and each is explicitly checkpointed at lifecycle transitions rather than on a fixed calendar, which is how MIAR embeds risk control into delivery rather than running it alongside it (Fig. 1).

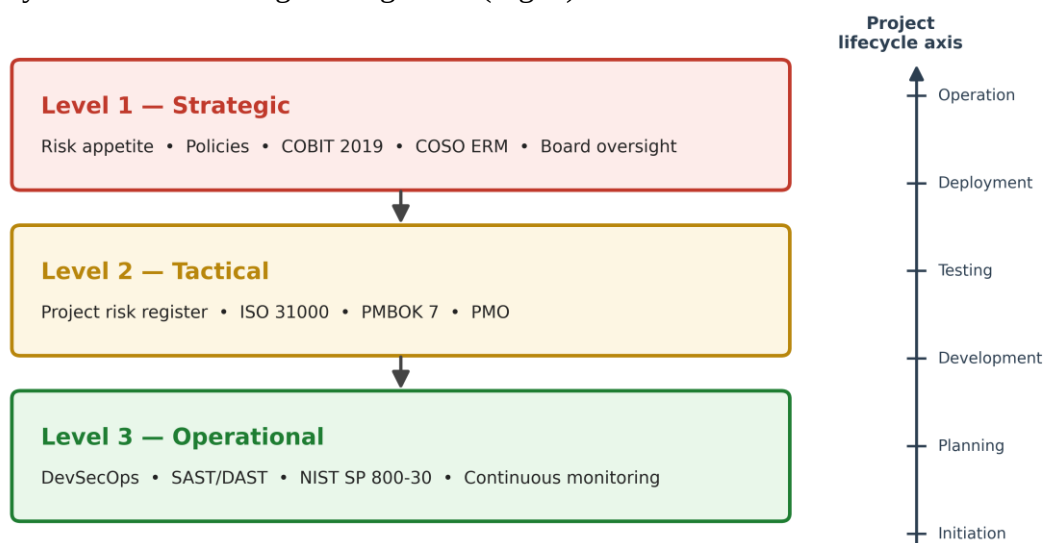


Fig. 1. Three-level Multi-Level Integrated Risk Architecture (MIAR).

An extended, seven-category IT risk taxonomy with 24 subcategories operationalizes MIAR (technical, organizational, financial, security, legal/compliance, external/vendor, and, introduced during the case validation, a Forensic Compliance category covering chain-of-custody and evidentiary-admissibility risks specific to law-enforcement-adjacent IT systems). The taxonomy defines, for each subcategory, typical triggers and early-warning indicators used both by human reviewers and as calibration examples for the AI module described in Section 2.5.

2.3 Risk Assessment Methodology

Two complementary assessment modes are used, following established quantitative-risk-modelling practice for IT projects [24]. Qualitative assessment uses a standard 5×5

probability–impact matrix: a group of at least three reviewers (project manager, tech lead, client representative) scores each risk independently; scores are consolidated by median rather than mean to reduce the influence of outliers, and disagreements exceeding two points trigger mandatory discussion with the rationale documented. Quantitative assessment uses Monte Carlo simulation: for each risk, a probability of occurrence and a PERT-distributed minimum/most-likely/maximum budget impact are defined; 10,000 iterations are run, in each of which occurrence is sampled and, if the risk occurs, an impact is drawn from its distribution; the resulting distribution of cumulative budget deviation yields P50 (expected loss), P80 (recommended contingency reserve) and P95 (conservative reserve) quantiles.

Four response strategies are used for negative risks, following ISO 31000 and PMBOK 7: Avoid (eliminate the risk by changing the plan, e.g., replacing an unstable third-party dependency), Transfer (move the financial consequence to a third party, e.g., insurance or SLA-backed outsourcing), Mitigate (reduce probability and/or impact, e.g., automated testing, code review, training), and Accept (a documented decision to take no action, either actively monitored or passively logged). Each risk of Medium level or above receives a written response plan (owner, deadline, concrete actions, escalation triggers, contingency plan, residual score), and an escalation matrix routes notification by risk level (Low: weekly, team; Medium: twice weekly, PM/tech lead; High: immediate, PM/sponsor; Critical: immediate, all channels).

2.4 The Risk Exposure Reduction Index (RERI) and Risk Management Index (RMI)

Existing effectiveness measures have structural limitations for cross-project comparison: Residual Risk ($\sum P_i \cdot I_i$) is an absolute quantity that is not comparable across projects of different scale; Risk Score Reduction ($RE_{\text{before}} - RE_{\text{after}}$) is an absolute delta that depends on the number of risks and the impact scale used; Risk Burn-down (the count of open risks over time) ignores severity; and the Risk Reduction Ratio ($RE_{\text{after}} / RE_{\text{before}}$) reports the residual share rather than the reduction share and is not directly interpretable. Table 1 summarizes these limitations.

Table 1. Existing risk-effectiveness metrics and their limitations relative to RERI.

Metric	Formula	Limitation
Residual Risk	$\sum P_i \cdot I_i$	Absolute value; not comparable across projects of different scale
Risk Score Reduction	$RE_{\text{before}} - RE_{\text{after}}$	Absolute delta; depends on risk count and impact scale
Risk Burn-down	Count of open risks over time	Ignores severity/weight; does not reflect net effectiveness
Risk Reduction Ratio	$RE_{\text{after}} / RE_{\text{before}}$	Reports the residual share, not the reduction share; unintuitive

RERI removes these limitations through normalization, criticality weighting and two correction factors. The base form is:

$$RERI_base = 1 - (RE_after / RE_before) \quad (1)$$

where $RE_before = \sum(P_i \cdot I_i)$ is total risk exposure before mitigation and $RE_after = \sum(P_i' \cdot I_i')$ is residual exposure after mitigation. $RERI_base \in (-\infty, 1]$: 0 indicates no effect, 1 indicates complete elimination, and negative values indicate that the applied measures were counterproductive. Because it is normalized by RE_before , $RERI_base = 0.641$ means “a 64.1% reduction” regardless of whether the project registry contains 5 or 500 risks, which the four metrics in Table 1 cannot express.

The extended form adds a criticality weight w_i , a cost factor and a timing factor:

$$RERI_ext = 1 - [\sum(w_i \cdot P_i' \cdot I_i') / \sum(w_i \cdot P_i \cdot I_i)] \cdot C_factor \cdot T_factor \quad (2)$$

with $w_i = 1.0$ for Low/Medium risks (score ≤ 9), 1.5 for High risks (score 10–14) and 2.0 for Critical risks (score 15–25), reflecting the empirical fact that losses from an unmitigated critical risk are disproportionately larger than losses from a medium one. The cost factor discounts effectiveness when mitigation is expensive relative to the value it protects:

$$C_factor = 1 - (C_mit / V_protected), \quad C_factor = 0 \text{ if } C_mit \geq V_protected \quad (3)$$

where C_mit is the total cost of implemented measures and $V_protected$ is the estimated value of avoided losses. The timing factor discounts effectiveness when mitigation occurs late in the lifecycle:

$$T_factor = 1 - \alpha \cdot (T_mit / T_project) \quad (4)$$

where T_mit is the point at which key measures were implemented, $T_project$ is total lifecycle duration, and $\alpha \in [0, 1]$ is a decay coefficient; $\alpha = 0.3$ was selected by calibration against three completed reference projects and produced a $RERI_ext$ variation of only ± 0.04 for $\alpha \in [0.2, 0.4]$, indicating low sensitivity to the exact value chosen. Because P_i and I_i are expert estimates and carry subjective uncertainty, a 95% confidence interval for RERI is obtained by bootstrap resampling: each P_i and I_i is modeled as a triangular distribution, $N = 1,000$ simulations are drawn, and RERI is recomputed for each draw to obtain $RERI = M \pm 1.96\sigma$.

The complementary Risk Management Index (RMI) tracks operational coverage rather than exposure reduction:

$$RMI = (active \text{ risks with an approved response plan} / total \text{ active risks}) \times 100\% \quad (5)$$

with $RMI \geq 80\%$ interpreted as an acceptable level of risk-management coverage, 50–79% as requiring attention, and below 50% as critical.

2.5 IT Risk Analyzer: System Architecture and the AI Module

Three candidate approaches to automated risk identification were compared before implementation: rule-based keyword matching, classical supervised ML classifiers (e.g.,

SVM, Random Forest), and LLM-based prompting. Rule-based systems are explainable but require continuous manual rule maintenance and generalize poorly to non-standard specifications. Classical ML requires a labeled dataset of several hundred examples per category, performs poorly on short ToR excerpts, and only classifies risks rather than generating mitigation plans. LLM-based prompting requires no labeled corpus, generates a structured risk register with descriptions, probabilities and mitigation plans, and adapts to new domains through the prompt alone, at the cost of dependence on an external API and non-deterministic output; the latter was mitigated with forced JSON output and a jsonrepair post-processing step. Given an academic development budget and the absence of a labeled Kazakhstani risk corpus, the LLM approach was selected; among candidate models, Claude Sonnet was chosen as primary (strongest structured-JSON output and instruction-following in internal comparison) with GPT-4o configured as an automatic fallback on rate-limit or server errors.

IT Risk Analyzer implements a three-tier architecture. The frontend is a React 18 / TypeScript single-page application that renders the risk register, the 5×5 heat-map, Monte Carlo charts and a Kanban board, and consumes a server-sent-events (SSE) stream for real-time analysis progress. The backend is a Node.js 20 / Express API server (19 route groups) that handles authentication (JWT, bcrypt-hashed passwords), document ingestion (DOCX/PDF/TXT), orchestration of AI requests, RERI/RMI computation, and external integrations (Jira, Azure DevOps, Slack). The AI module wraps the Claude API (model claude-sonnet-4-6, max_tokens = 8192, temperature = 0.3 for analysis) with a system prompt that fixes the model's role and output schema, uses chain-of-thought reasoning to iterate systematically over risk categories, and applies few-shot calibration examples per category; the input specification text is treated strictly as user content, which together with the fixed system role mitigates prompt-injection risk, consistent with general information-security risk-management guidance [25]. Data are persisted in PostgreSQL 16 (14 tables, including generated columns for risk score and level to guarantee consistency without triggers) with Redis used for caching and rate-limit queuing (Fig. 2). The platform is multi-tenant (organizations, roles, subscription tiers) and exposes a developer API.

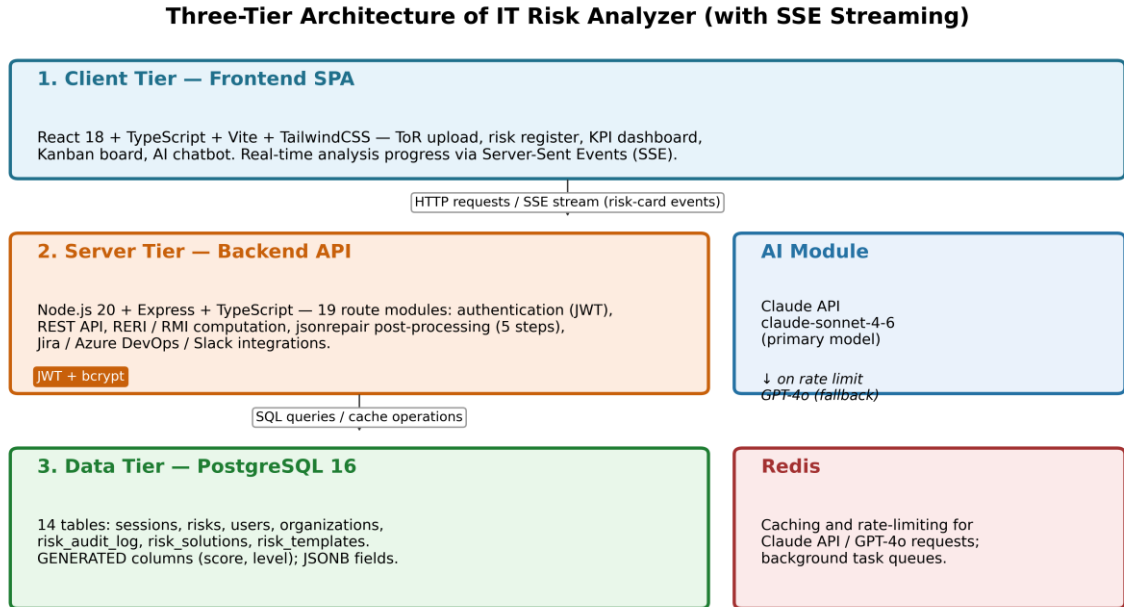


Fig. 2. Conceptual (C4 Level 1) architecture of the IT Risk Analyzer web service.

2.6 Evaluation Design

The AI module was evaluated on five real technical specifications from distinct domains (e-commerce platform, banking mobile application, manufacturing ERP system, an educational AI chatbot, and a government API-integration project), collected between October and November 2025. For each specification, the author, acting as an experienced project manager / systems analyst, produced an independent expert risk list following PMBOK 7 before reviewing the AI module's output; this list was then cross-verified against the NIST SP 800-30, OWASP Top 10 and PMBOK 7 taxonomies, and only risks confirmed by at least one published catalogue were retained as ground truth, which reduces confirmation-bias risk inherent in using the methodology's own author as the sole rater. An AI-identified risk was scored as a true positive if its category matched or closely corresponded to a ground-truth risk and its stated cause/consequence addressed the same underlying threat; this matching criterion was fixed before comparison began (pre-registered).

Precision, recall and F1 were computed as:

$$\text{Precision} = TP / (TP + FP); \text{Recall} = TP / (TP + FN); F1 = 2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall}) \quad (6)$$

where TP is a risk found by the AI module and confirmed by the expert ground truth, FP is an AI-identified risk not confirmed, and FN is a ground-truth risk missed by the AI

module. The same case project described in Section 2.1 was used to validate RERI, RMI and ROI empirically, with expected avoided losses (EL) computed as $EL = P \cdot \text{Impact_financial}$ per critical/high risk, drawing on contractual penalty clauses, documented legal precedent for evidentiary admissibility, and historical infrastructure cost data, and methodology ROI computed as $(EL_avoided - C_mitigation) / C_mitigation \times 100\%$.

3. Results

3.1 AI Module Risk-Identification Quality

Table 2 reports precision, recall and F1 for the AI module across the five evaluated technical specifications, together with the underlying true-positive, false-positive and false-negative counts.

Table 2. AI-module risk-identification quality across five real technical specifications.

Specification	GT	AI found	TP	FP	FN	Precision	Recall	F1
E-commerce platform	18	22	16	6	2	72.7%	88.9%	80.0%
Banking mobile app	23	24	20	4	3	83.3%	87.0%	85.1%
Manufacturing ERP	27	25	21	4	6	84.0%	77.8%	80.8%
Educational AI chatbot	15	19	14	5	1	73.7%	93.3%	82.4%
Government API integration	20	22	17	5	3	77.3%	85.0%	81.0%
Mean	20.6	22.4	17.6	4.8	3.0	78.2%	86.4%	81.9%

Figure 3 presents the same precision/recall/F1 results graphically (Fig. 3).

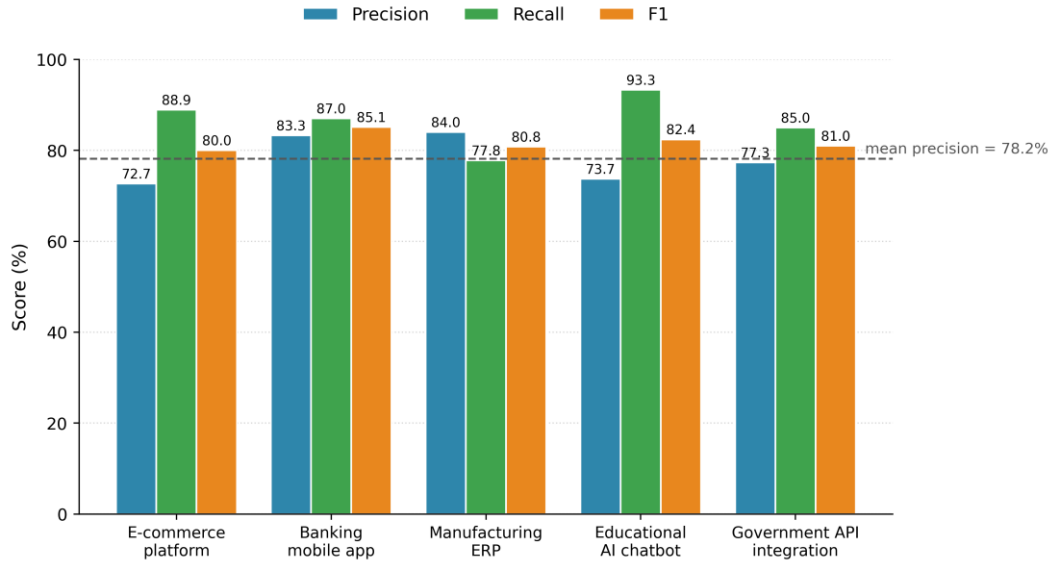


Fig. 3. AI-module precision, recall and F1 across five real technical specifications.

Result 1 (Table 2) demonstrates Contribution 3: the AI module reaches a mean F1 of 81.9% (precision 78.2%, recall 86.4%), exceeding the 78% threshold set for hypothesis H1 and confirming applicability to first-pass risk screening across heterogeneous project domains. The AI module systematically identified on average 1.8 more risks than the expert baseline per specification ($FP > FN$), which is an acceptable asymmetry for a screening tool: a false alarm is materially cheaper to discard than a missed risk. Mean generation time was 18–25 s per specification (median time-to-first-token 1.2–1.8 s), with 19–27 risks produced per registry, and the underlying service maintained 99.2% uptime over the 45-day measurement window (52 minutes of downtime out of 1,080 hours, attributable to planned maintenance).

3.2 Case-Study Validation: NIC CyberTech

Table 3 shows the RERI_ext computation for the case project, aggregating the three Critical risks, two High risks explicitly broken out, and the remaining 25 Medium/Low risks.

Table 3. Weighted risk-exposure reduction (RERI_ext components) for the case project.

(RE_b / w·RE_b: exposure and weighted exposure before mitigation; RE_a / w·RE_a: after mitigation.)

Risk	Level	w_i	RE_b	w·RE_b	RE_a	w·RE_a	Reduction
Chain-of-custody (evidentiary integrity)	Critical	2.0	15	30.0	5	10.0	67%

Video-analysis performance degradation	Critical	2.0	16	32.0	4	8.0	75%
ML false-positive rate (multilingual)	Critical	2.0	15	30.0	3	6.0	80%
Forensic-compliance (legal admissibility)	High	1.5	12	18.0	4	6.0	67%
ML dependence on external API	High	1.5	10	15.0	4	6.0	60%
Remaining 25 Medium/Low risks	Medium	1.0	127	127.0	50	50.0	61%
Weighted total	—	—	—	252.0	—	86.0	—

With $C_factor = 1 - (6,200,000 / 43,500,000) = 0.857$ and $T_factor = 1 - 0.3 \cdot 0.4 = 0.88$ (key measures implemented within the first four of ten active-development months), Equation (2) gives $RERI_ext = 1 - (86.0/252.0) \cdot 0.857 \cdot 0.88 = 0.743$. For the base (unweighted) formulation, $RE_before = 195$ and $RE_after = 70$ across all 30 risks, giving $RERI_base = 1 - (70/195) = 0.641$. A bootstrap procedure ($N = 1,000$, ± 1 point perturbation on P and I) produced $RERI_base = 0.641 \pm 0.078$, 95% CI [0.563, 0.719]; even the pessimistic bound of this interval remains above the 0.6 threshold associated with “high effectiveness” (Fig. 4).

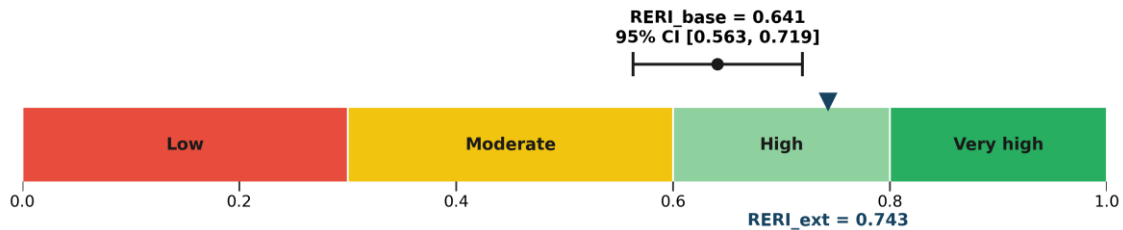


Fig. 4. $RERI_base$ and $RERI_ext$ on the effectiveness scale, with the 95% bootstrap confidence interval for $RERI_base$.

Result 2 (Table 3) demonstrates Contribution 4: of the 30 registered risks, 27 (all Critical, all High, and all Medium risks) carried an approved response plan, giving $RMI = 27/30 = 90\%$, above the 80% threshold set for hypothesis H2. Expected avoided losses were estimated at ≈ 43.5 million KZT (legal risk from inadmissible evidence, contractual penalty for delay, and GPU-infrastructure cost overrun), against a mitigation cost of 6.2 million KZT, yielding a methodology ROI of $(43,500,000 - 6,200,000) / 6,200,000 \times 100\% \approx 601\%$.

Table 4 summarizes the verification status of the four sub-hypotheses and the rejected null hypothesis against their pre-registered thresholds.

Table 4. Hypothesis verification against pre-registered thresholds.

Hypothesis	Threshold	Observed result	Status
H1 — LLM identification quality	$F1 \geq 78\%$	$F1 = 81.9\%$, Recall = 86.4%	Confirmed
H2 — operational coverage (RMI)	$RMI \geq 80\%$	$RMI = 90\%$	Confirmed
H3 — RERI bootstrap stability	$95\% \text{ CI width} \leq 0.20$	$CI = [0.563, 0.719]$, width 0.156	Confirmed
H4 — applied exposure reduction	$RERI_base \geq 0.5$	$RERI_base = 0.641$	Confirmed
H0 — no significant advantage	$RERI < 0.5$	0.641, CI lower bound 0.563	Rejected

Result 3 (Table 4) demonstrates Contribution 1 and Contribution 2 jointly: the integrated MIAR architecture combined with the RERI/RMI measurement layer produced a risk-exposure reduction that clears its pre-registered threshold with a practically significant margin (+28 percentage points over the H4 threshold), and the reduction is stable under resampling (H3).

4. Discussion

RQ1 asked whether a single integrated architecture could remove the structural gaps between ISO 31000, PMBOK 7, COBIT 2019 and NIST SP 800-30. The results presented in Section 3.2 demonstrate that the observed RERI of 0.641 was achieved not by applying any one standard more rigorously, but through the joint operation of MIAR's three levels: corporate governance (COBIT 2019/COSO ERM), tactical planning (ISO 31000/PMBOK 7) and operational control (NIST SP 800-30/OWASP). This supports an affirmative answer to RQ1 for the government-project context studied and is consistent with the broader observation in the software-risk literature that isolated application of a single standard leaves significant blind spots [10, 11].

RQ2 asked whether current LLMs can identify IT risks from specification text at expert-comparable accuracy ($F1 \geq 75\%$). The AI module reached $F1 = 81.9\%$ (Table 2), exceeding this threshold and comparable to published international results: Wang & Liu [14] report $F1 \approx 81\%$ for an ensemble LLM/expert-rules approach, and Kaur et al. [12] report 74% for a RAG/NLP-rules approach, while requiring no labeled training corpus, which is the principal advantage of the prompting approach over Wang & Liu's ensemble method for a setting, such as Kazakhstan's, where no domain-specific risk dataset exists. The main source of false negatives was the specialized Forensic Compliance category, a domain essentially absent from public LLM training distributions; this indicates a boundary condition rather than a failure of the approach: LLMs identify standard risk categories (OWASP, PMBOK, ISO 31000) reliably and specialized, narrow-domain risks less reliably. Consistent with this boundary, 8 of the 30 risks in the case project (27%), including four of the eight High/Critical risks ultimately identified, were added by the human team after AI-

assisted screening rather than by the model itself. This positions the AI module as a first-pass screening instrument that materially accelerates identification without removing the need for expert verification, rather than as a replacement for it.

RQ3 asked whether the strategy produces measurable, verifiable exposure reduction ($RERI \geq 0.5$, $RMI \geq 80\%$). Table 4 shows both thresholds were exceeded, with the RERI result stable under bootstrap resampling (95% CI [0.563, 0.719]). The high ROI (601%) observed in this case is attributable to project structure rather than to a methodological artifact: three critical risks with combined expected loss of ≈ 38 million KZT belonged to legal and forensic-compliance categories in which realization risked termination of a 186-million-KZT government contract, while mitigation cost only 3.3% of the contract value. For projects with less catastrophic risk profiles, typical commercial products, a substantially lower ROI (an estimated 50–200%) would be expected, and this should temper generalization of the 601% figure specifically, even though the underlying RERI/RMI improvements are expected to transfer.

Compared with prior Kazakhstani work, the rule-based expert-audit approach of Dilmukhanbetov [18] reported approximately 68% accuracy on a limited sample without an automation component, and Zhaksybekov [19] applied a matrix-based information-security risk assessment without automated identification; the present study improves on identification accuracy by roughly 14 percentage points and additionally provides an integrated architecture, a normalized effectiveness index, and a documented production validation with a signed implementation act, a combination not previously reported in the regional literature reviewed.

This study has limitations. The validation rests on a single case project with the author acting simultaneously as researcher, systems analyst and rater, which is a source of potential bias, partially mitigated by pre-registering the scoring criterion and by cross-verifying ground truth against three independent published catalogues (NIST SP 800-30, OWASP Top 10, PMBOK 7). The AI-module evaluation sample of five specifications was limited by the availability of real, de-identifiable technical specifications; expanding this sample to 20 or more specifications is a priority for follow-up work. RERI does not model correlation between risks (two dependent risks are currently treated as independent) and its timing factor relies on an empirically calibrated but ultimately judgment-based decay coefficient α . Finally, generalization beyond government forensic-compliance projects to typical commercial IT delivery requires replication on three to five additional projects, which the authors identify as the most important next step.

5. Conclusion

This paper set out to close a gap between mature but fragmented international risk-governance standards and the practical needs of IT delivery teams, particularly in a market, Kazakhstan, where integrated tooling and measurable risk-management outcomes remain scarce. The proposed Multi-Level Integrated Risk Architecture (MIAR) generalizes beyond the specific numbers reported here: by construction, it embeds governance, tactical planning

and operational control into any IT project lifecycle rather than treating risk management as a parallel, periodic activity, and it is this structural property, not any single standard applied in isolation, that the case results attribute the observed exposure reduction to.

The scientific novelty claimed in Section 1, an integrated cross-standard architecture (MIAR), a normalized and criticality-weighted effectiveness index (RERI) with bootstrap-quantified uncertainty, and a verified LLM-based prompting framework for risk identification, has been substantiated by the results reported in Section 3: an F1 of 81.9% for automated identification, comparable to international benchmarks without requiring a labeled corpus; a RERI_base of 0.641 with a 95% confidence interval that remains above the high-effectiveness threshold even at its lower bound; and a documented, independently signed implementation record on a real government IT project. Where the evidence is necessarily partial (a single validated case, a five-specification AI-evaluation sample), this has been stated explicitly rather than implied away, and the ROI figure of 601% in particular should be read as a property of this project's risk structure rather than a generic expectation for all IT projects.

Beyond the specific case, the ambiguous data point worth flagging is the boundary behavior of the LLM component: it screens standard risk categories well and narrow, specialized domains (here, forensic compliance) less reliably, which argues for treating LLM-based identification as an accelerant for expert review rather than a substitute for it in any deployment. For future research, three directions follow directly from the limitations identified above: fine-tuning a smaller open-source model on an organization's accumulated risk-register data to reduce dependence on an external API and enable on-premises deployment for sensitive projects; extending the taxonomy's Forensic Compliance category to adjacent law-enforcement IT domains with formal certification against Kazakhstani regulatory requirements; and replicating the quasi-experimental validation on several commercial (non-government) projects to establish how far the reported RERI/RMI/ROI figures generalize outside a high-stakes public-sector setting. Taken together, the results indicate that pairing an integrated governance architecture with LLM-assisted identification is a practically deployable, measurably effective strategy for IT risk management, and provide a reproducible basis for extending that claim to a broader population of projects.

Acknowledgments

The author thanks the management and project team of the case organization (referred to herein under the pseudonym “NIC CyberTech”) for their cooperation in the validation study and for authorizing the de-identified disclosure of project outcomes. No external funding was received for this research.

Author Contributions

Conceptualization: F.K.; Methodology: F.K.; Software: F.K.; Validation: F.K.; Formal analysis: F.K.; Investigation: F.K.; Data curation: F.K.; Writing — original draft preparation: F.K.; Writing — review and editing: F.K.; Visualization: F.K.

Conflicts of Interest

The author declares no conflict of interest.

Data Availability Statement

The de-identified risk register and RERI computation underlying Section 3.2 are available from the author on reasonable request; the original technical specification and full project records are not publicly available owing to contractual confidentiality obligations of the client organization. The source code of the IT Risk Analyzer platform is not currently open-sourced.

References

- [1] Standish Group. (2023). Chaos Report 2023: Beyond Infinity. Standish Group International.
- [2] Ministry of Digital Development, Innovation and Aerospace Industry of the Republic of Kazakhstan. (2024). Digital Development Report of Kazakhstan 2024. MDDIA.
- [3] International Organization for Standardization. (2018). ISO 31000:2018 — Risk management — Guidelines. ISO.
- [4] Project Management Institute. (2021). A Guide to the Project Management Body of Knowledge (PMBOK Guide) (7th ed.). PMI.
- [5] National Institute of Standards and Technology. (2012). NIST SP 800-30 Rev. 1 — Guide for Conducting Risk Assessments. NIST.
- [6] ISACA. (2019). COBIT 2019 Framework: Governance and Management Objectives. ISACA.
- [7] Committee of Sponsoring Organizations of the Treadway Commission. (2017). Enterprise Risk Management — Integrating with Strategy and Performance. COSO.
- [8] Hubbard, D. W. (2020). The Failure of Risk Management: Why It's Broken and How to Fix It (2nd ed.). Wiley.
- [9] Brooks, F. P. (1995). The Mythical Man-Month: Essays on Software Engineering (Anniversary ed.). Addison-Wesley.
- [10] DeMarco, T., & Lister, T. (2003). Waltzing with Bears: Managing Risk on Software Projects. Dorset House.
- [11] Humble, J., & Farley, D. (2010). Continuous Delivery: Reliable Software Releases through Build, Test, and Deployment Automation. Addison-Wesley.
- [12] Kaur, P., Singh, A., & Sharma, M. (2023). AI-driven risk classification in cybersecurity: A RAG-based approach. *Journal of Information Security and Applications*, 75, 103421.

- [13] Mates, R., & Hobbs, C. (2022). GPT-3 for project risk identification: Accuracy and limitations. *Project Management Journal*, 53(4), 412–425.
- [14] Wang, L., & Liu, H. (2024). Ensemble approach for IT risk identification using LLM and expert rules. *IEEE Transactions on Software Engineering*, 50(3), 1124–1138.
- [15] Chen, X., Zhang, Y., & Li, W. (2025). Applying Claude 3 for automated risk register generation in startup projects. *Proceedings of ICSE 2025* (pp. 344–353). IEEE.
- [16] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- [17] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Neelakantan, A. et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- [18] Dilmukhanbetov, A.K. (2022). Upravlenie IT-riskami v gosudarstvennykh proektakh Kazakhstana: problemy i resheniya [IT risk management in Kazakhstani government projects: problems and solutions]. *Vestnik KazNU. Seriya matematiki, mekhaniki i informatiki*, 4(116), 112–120.
- [19] Zhaksybekov, R.N. (2023). Metodika kolichestvennoy otsenki riskov informatsionnoy bezopasnosti dlya organizatsiy gosudarstvennogo sektora RK [A method for quantitative information-security risk assessment for public-sector organizations in Kazakhstan] *Problemy Informatiki*, 1(58), 43–52.
- [20] Buribayeva, A.K. (2024). Aktual'nye problemy upravleniya riskami v tsifrovoy transformatsii gosudarstvennykh organov Kazakhstana [Current problems of risk management in the digital transformation of Kazakhstani government agencies]. *Vestnik KazUTB. Seriya Informatsionnye Tekhnologii*, 2, 55–64.
- [21] Khassanov, F. F., & Shaikhanova, A. K. (2025). Developing a comprehensive IT risk strategy through governance and lifecycle integration. *Central Asian Journal of Information Technology*, 3(2), 18–29.
- [22]. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105.
- [23] OWASP Foundation. (2021). OWASP Top Ten 2021. Open Web Application Security Project.
- [24] Zhao, J., & Chen, R. (2023). Quantitative risk modelling in IT projects using Monte Carlo simulation. *International Journal of Project Management*, 41(5), 523–536.
- [25] International Organization for Standardization. (2022). ISO/IEC 27005:2022 — Information security, cybersecurity and privacy protection — Guidance on managing information security risks. ISO.

Aisagaliyev Islam

Bachelor Student, School of Cybersecurity

241261@astanait.edu.kz

Astana IT University, Kazakhstan

Albara Malkawi

Business Developer

albara@mihsad.com

Mihsad – Business Development, KSA, Jordan

SECURING THE PHYSICAL-DIGITAL FRONTIER: AI FOR ANOMALY DETECTION IN OT, IoT, AND CRITICAL INFRASTRUCTURE

Abstract: Operational Technology (OT) networks, heterogeneous Internet-of-Things (IoT) deployments, and the information-technology systems that increasingly converge with them form a physical-digital frontier whose security properties differ markedly from conventional enterprise IT, since legacy industrial protocols, decade-long device lifecycles, and strict latency constraints rule out many controls that work well in cloud and corporate environments while the consequence of compromise extends from data loss to physical harm. This study reviews the use of artificial intelligence, specifically agentic, large language model (LLM)-augmented systems, for anomaly detection across this frontier, drawing on a comparative analysis of current detection literature, field-deployment reports from energy utilities, and the detection requirements that distinguish OT, IoT, and IT layers. The review separates cognitive support, meaning the interpretation of noisy multi-source telemetry, from execution support, meaning the automated isolation of compromised nodes, and formalizes a defensive operational uplift metric $U = T_{\text{manual}} / T_{\text{AI}}$ to quantify how much AI assistance compresses the time between an anomaly occurring and a validated, actionable alert. Reported field results show uplift ranging from 1.2-2x for novel anomaly classes with no prior context to 4-10x for cross-source correlation and alert enrichment, with transformer-based and LLM-correlation techniques consistently outperforming signature-based and purely behavioral baselines while preserving the explainability that safety-critical OT deployments require. The study concludes that the salient contribution of AI in this domain is workflow acceleration rather than autonomy, that edge deployment of quantized models is the dominant pattern where latency or data sovereignty rules out cloud inference, and that defenders must anticipate a symmetric risk in which attackers use the same agentic techniques to evade detection. These findings inform how critical-infrastructure operators can prioritize AI investment toward analyst support rather than unsupervised automation.

Keywords: *agentic AI; anomaly detection; critical infrastructure; defensive uplift; edge AI; industrial control systems; IoT security; LLM-based monitoring; operational technology; explainable AI*

Обеспечение безопасности на границе физического и цифрового миров: ИИ для обнаружения аномалий в операционных технологиях, Интернете вещей и критической инфраструктуре.

Аннотация: Сети операционных технологий (OT), гетерогенные среды Интернета вещей (IoT) и все более сближающиеся с ними информационные системы (IT) формируют физико-цифровое пространство. Свойства безопасности этого пространства заметно отличаются от традиционных корпоративных IT-систем. Устаревшие промышленные протоколы, десятилетние жизненные циклы устройств и строгие ограничения по задержке исключают использование многих мер контроля, которые эффективно работают в облачных и корпоративных средах. При этом последствия компрометации здесь варьируются от потери данных до физического ущерба.

В данном исследовании рассматривается использование искусственного интеллекта, а именно агентных систем на базе больших языковых моделей (LLM), для обнаружения аномалий в этом пространстве. Анализ основан на сравнении современной литературы по детекции, отчетов о практическом развертывании на энергетических предприятиях, а также требований к обнаружению угроз, которые отличают уровни OT, IoT и IT. В обзоре разделяются когнитивная поддержка (интерпретация зашумленной телеметрии из множества источников) и исполнительная поддержка (автоматическая изоляция скомпрометированных узлов). Также формализуется метрика повышения эффективности защиты $U = T_{\text{manual}} / T_{\text{AI}}$ для количественной оценки того, насколько помощь ИИ сокращает время между возникновением аномалии и созданием проверенного, готового к действию оповещения.

Результаты практического применения показывают рост эффективности в пределах 1.2–2х для новых классов аномалий без предварительного контекста и в 4–10х для межсистемной корреляции и обогащения данных об оповещениях. При этом методы на основе трансформеров и LLM-корреляции стабильно превосходят сигнатурные и чисто поведенческие базовые показатели, сохраняя объяснимость, необходимую для критически важных с точки зрения безопасности систем OT. В исследовании делается вывод, что главным преимуществом ИИ в этой области является ускорение рабочих процессов, а не автономность. При этом граничное развертывание (edge deployment) квантованных моделей выступает доминирующим паттерном там, где задержка или суверенитет данных исключают облачные вычисления. Наконец, защитникам необходимо учитывать симметричный риск, при котором злоумышленники используют те же агентные методы для обхода систем обнаружения. Эти выводы помогают операторам критической инфраструктуры правильно расставлять приоритеты при инвестировании в ИИ, ориентируясь на поддержку аналитиков, а не на неконтролируемую автоматизацию.

Ключевые слова: агентный ИИ; обнаружение аномалий; критическая инфраструктура; повышение эффективности защиты; граничный ИИ; промышленные системы управления; безопасность IoT; мониторинг на базе LLM; операционные технологии; объяснимый ИИ.

Физикалық-сандық шекараны қорғау: OT, Заттар интернеті және маңызды инфрақұрылымдағы аномалияларды анықтауға арналған жасанды интеллект

Аңдатпа: Операциялық технологиялар (OT) желілері, гетерогенді заттар интернеті (IoT) жүйелері және олармен барған сайын кірігіп жатқан ақпараттық технологиялар (IT) жүйелері физикалық-цифрлық кеңістікті құрайды. Бұл кеңістіктің қауіпсіздік қасиеттері дәстүрлі корпоративтік IT-жүйелерден айтарлықтай ерекшеленеді. Ескірген өнеркәсіптік хаттамалар, құрылғылардың ондаған жылдарға созылатын өмірлік циклі және кідіріс уақытына қойылатын қатаң шектеулер бұлтты және корпоративтік ортада тиімді жұмыс істейтін көптеген бақылау шараларын қолдануға мүмкіндік бермейді. Сонымен қатар, бұл жүйелердің бұзылу салдары деректерді жоғалтудан бастап физикалық зиян келтіруге дейін жетуі мүмкін.

Бұл зерттеуде аталған кеңістіктегі аномалияларды анықтау үшін жасанды интеллекттің, атап айтқанда үлкен тілдік модельдермен (LLM) күшейтілген агенттік жүйелердің қолданылуы қарастырылады. Талдау қазіргі анықтау әдебиеттерін салыстырмалы бағалауға, энергетикалық кәсіпорындардың практикалық енгізу есептеріне және OT, IoT және IT деңгейлерін ерекшелейтін қауіптерді анықтау талаптарына негізделген. Шолуда когнитивтік қолдау (көптеген дереккөздерден алынған шулы телеметрияны түсіндіру) және атқарушылық қолдау (бұзылған түйіндерді автоматты түрде оқшаулау) бөлек қарастырылады. Сондай-ақ, ЖИ көмегі аномалияның пайда болуы мен тексерілген, әрекет етуге дайын ескертудің жасалуы арасындағы уақытты қаншалықты қысқартатынын сандық бағалау үшін қорғаныстың тиімділігін арттыру метрикасы $U = T_{\text{manual}} / T_{\text{AI}}$ форматталады.

Практикалық қолданыс нәтижелері бұрын контексті болмаған жаңа аномалия кластары үшін тиімділіктің 1.2–2 есе, ал жүйеаралық корреляция мен ескертулерді деректермен байыту үшін 4–10 есе артқанын көрсетеді. Бұл ретте трансформерлер мен LLM-корреляциясына негізделген әдістер қауіпсіздік тұрғысынан маңызды OT жүйелеріне қажетті түсіндірмелілікті сақтай отырып, сигнатуралық және таза мінез-құлықтық базалық көрсеткіштерден үнемі асып түседі. Зерттеу бұл саладағы ЖИ-дің негізгі үлесі автономия емес, жұмыс процестерін жеделдету деген қорытындыға келеді. Сондай-ақ, кідіріс немесе деректер егемендігі бұлтты есептеулерді қолдануға мүмкіндік бермейтін жерлерде квантталған модельдерді шеткі құрылғыларда енгізу (edge deployment) негізгі үлгі болып табылады. Сонымен қатар, қорғаушылар қаскөйлердің де жүйені айналып өту үшін дәл осындай агенттік әдістерді қолдануы

мүмкін симметриялы тәуекелді ескеруі керек. Бұл тұжырымдар критикалық инфрақұрылым операторларына ЖИ-ге инвестиция салу кезінде бақылаусыз автоматтандыруға емес, сарапшыларды қолдауға басымдық беруге көмектеседі.

Түйін сөздер: агенттік ЖИ; аномалияларды анықтау; критикалық инфрақұрылым; қорғанысты тиімділендіру; шеткі ЖИ (edge AI); өнеркәсіптік басқару жүйелері; IoT қауіпсіздігі; LLM негізіндегі мониторинг; операциялық технологиялар; түсіндірмелі ЖИ.

1. Introduction (Literature Review)

Critical infrastructure, including electricity, water, transportation, healthcare delivery, and manufacturing, has spent the past two decades digitizing the link between sensors and decisions. Programmable Logic Controllers (PLCs) talk to historians; Remote Terminal Units (RTUs) report to SCADA over Ethernet rather than serial; building-management IoT devices share networks with corporate laptops. The security properties of these blended environments are not those of traditional IT. Devices were designed for availability and safety, not confidentiality; firmware lifecycles are measured in years, sometimes decades; protocol stacks such as Modbus, DNP3, IEC 60870-5-104, S7, and BACnet authenticate poorly or not at all [1], [2]. The 2010 Stuxnet incident, the 2015 and 2016 Ukrainian grid attacks, the 2021 Colonial Pipeline shutdown, and the 2024 multi-utility wiper campaigns have made clear that compromise in this domain produces physical, not merely informational, consequences [3], [4].

The defensive challenge is structural. Attackers need only one usable path, such as a single dropped packet that an intrusion-detection system misclassifies, one PLC firmware write that the historian does not flag, or one engineering workstation that an analyst clears too quickly. Defenders must process the full noise stream from thousands of sensors and tens of thousands of network flows, and they must do so under latency budgets that, in OT, are measured in milliseconds rather than seconds [5]. This defensive asymmetry mirrors the asymmetric advantage discussed in the offensive-AI literature [6]: the attacker's economics improve whenever AI lowers the cost of producing one useful output, while the defender's economics improve only when AI lowers the cost of correctly classifying every input. Because these two conditions are not symmetric, a naive transfer of offensive-AI patterns onto the defensive side tends to disappoint in practice.

Literature Review. Detection research for OT and IoT environments has evolved along three broad lines. The first line, signature-based and rule-based detection, relies on predefined protocol dissectors and blacklists; it remains valued for being deterministic and auditable but is structurally unable to recognize novel attacks or the absence of legitimate behavior [7], [8]. The second line, behavioral and statistical detection using autoencoders, recurrent networks, graph neural networks, and hidden-Markov models, learns a baseline of normal operation and flags deviation, which captures stealthy process-variable manipulation that signatures miss but produces high false-positive rates whenever legitimate process

changes resemble incipient attack [9], [10], [11]. The third and most recent line introduces agentic, LLM-augmented systems that unify heterogeneous telemetry, including PLC variable streams, network flows, gateway syslogs, and free-text operator notes, into a single interpretive narrative, drawing on the broader agentic-AI literature on reasoning-and-acting architectures and tool use [12], [13], [14], [15]. Across this evolution, a consistent finding is the analyst bottleneck: critical-infrastructure security operations centers (SOCs) are constrained less by detection capability than by the rate at which alerts can be validated and acted upon, since a single high-confidence alert on a turbine-control network may require coordination across several teams and a documented decision before any containment action [16].

Research Gap. Existing comparative work on AI-based intrusion and anomaly detection in OT and IoT settings concentrates heavily on detection accuracy reported in isolated benchmark studies, with comparatively little attention paid to how that accuracy translates into operational time savings for the human analyst who must still validate the alert, or to how detection requirements and human-in-the-loop constraints differ across the OT, IoT, and IT layers within the same organization. Studies that propose a quantitative metric for the defensive value of AI assistance are scarce, and almost none address the symmetric risk that the same agentic techniques deployed for defense can be repurposed by attackers to evade detection.

Research Question(s). This study addresses the following research question: how should the defensive contribution of agentic, LLM-augmented AI to OT, IoT, and critical-infrastructure anomaly detection be characterized and measured? Specifically, the paper examines: (1) how AI-driven detection techniques differ in effectiveness and deployment requirements across the OT, IoT, and IT layers; (2) how the defensive contribution of an AI agent can be quantified in a way that is operationally meaningful to a security operations center; and (3) what limitations and symmetric risks remain when AI is deployed for defense in this domain.

Scientific Novelty. The scientific novelty of this work consists in adapting the operational-uplift formula from the offensive-AI literature to the defensive context of OT and IoT anomaly detection, formalized as $U = T_{\text{manual}} / T_{\text{AI}}$, and in using this metric to organize a comparative analysis of detection techniques across the OT, IoT, and IT layers. This distinguishes the review from prior work in that it evaluates AI's contribution primarily in terms of analyst workflow acceleration rather than detection accuracy alone, and it explicitly maps the cognitive-support and execution-support roles of an AI agent onto the differing human-in-the-loop requirements of each layer.

Contributions. The main contributions of this study are:

- A review of AI-driven anomaly detection techniques for OT, IoT, and critical-infrastructure environments, organized around the cognitive-support and execution-support distinction;

- A formalization of the defensive operational uplift metric $U = T_{\text{manual}} / T_{\text{AI}}$, together with reported uplift ranges drawn from field deployments at energy utilities;
- A comparative analysis of detection requirements, noise characteristics, latency budgets, and human-in-the-loop necessity across the OT, IoT, and IT layers;
- Identification of the symmetric risk that AI-driven evasion poses to AI-driven detection, together with operational recommendations for critical-infrastructure operators.

The remainder of the paper is organized as follows: Section 2 presents the methods and materials underlying this review; Section 3 reports the comparative results across detection techniques and infrastructure layers; Section 4 discusses the findings and answers the research questions; Section 5 concludes the paper.

2. Methods and Materials

This study employs a comparative literature review methodology, supplemented by field-deployment data reported in industry and academic sources, to evaluate the use of agentic AI for anomaly detection across OT, IoT, and IT layers of critical infrastructure. The review draws on peer-reviewed journal articles, conference proceedings published between 2021 and 2026, standards documents (NIST SP 800-82, ISA/IEC 62443, IEC 61508), and industry threat-landscape and field-study reports from organizations including ENISA, MITRE, Dragos, and Microsoft Threat Intelligence.

The reviewed literature was classified into four categories of detection technique, namely signature-based systems, behavioral and statistical machine learning approaches, deep learning architectures, and agentic LLM-augmented systems. For each category, the review extracted reported detection signal, deployment constraints, explainability characteristics, and, where available, a quantitative defensive operational uplift value $U = T_{\text{manual}} / T_{\text{AI}}$, defined as the ratio of the wall-clock time required for an unaided analyst to move from raw telemetry to a validated, actionable alert (T_{manual}) to the corresponding time when supported by an AI agent that pre-correlates, hypothesizes, and enriches the alert (T_{AI}).

2.1 Cognitive and Execution Support

The review adopts the cognitive-support / execution-support decomposition established in the agentic-AI literature [12], [13] and applies it to the OT/IoT setting. Cognitive support refers to agent functions whose output is interpretation rather than action, such as synthesizing a hypothesis from a pressure-variable drift, an out-of-hours login, and a firmware-upload event on the same VLAN. Execution support refers to functions whose output materially changes the environment, such as isolating a node or throttling a control-loop setpoint, and carries a sharper risk profile because erroneous execution can itself produce a safety incident [17].

2.2 Comparative Criteria Across Infrastructure Layers

For the cross-layer comparison reported in Section 3, the review evaluated each of the OT, IoT, and IT layers against five criteria: data noise characteristics, latency budget, the

primary benefit that AI contributes, the necessity of human-in-the-loop (HIL) review, and the typical deployment pattern (edge-only, mixed edge/cloud, or cloud-native).

3. Results

The reviewed literature demonstrates a clear evolution from signature-based detection toward behavioral machine learning, deep learning, and most recently agentic LLM-augmented systems. Table 1 summarizes the comparative properties of AI-driven anomaly detection across the OT, IoT, and IT layers.

Table 1. Comparative properties of AI-driven anomaly detection across OT, IoT, and IT layers (HIL = human-in-the-loop).

Property	OT	IoT	IT
Data noise	Low-medium; deterministic process signals	High; heterogeneous, intermittent connectivity	High; user-driven traffic, broad legitimate variance
Latency budget	Strict; ms-level for control loops	Mixed; real-time to batch by asset	Relaxed; seconds to minutes acceptable
Primary AI benefit	Context fusion of process, network, and operator events	Behavioral baselining across device classes	Alert enrichment, correlation, NL triage
HIL necessity	Very high; safety-critical actions need authorization	Moderate; varies by criticality and reversibility	Low-moderate; many actions safely automatable
Deployment	Edge-only inference; no cloud egress	Mixed edge/cloud; device-compute bound	Cloud-native acceptable

Result 1 (Table 1) demonstrates Contribution 3 — the comparative analysis confirms that OT, IoT, and IT layers impose substantially different latency budgets and human-in-the-loop requirements, with OT requiring edge-only inference and very high HIL necessity while IT can tolerate cloud-native deployment and lower-touch automation.

Table 2 reports AI detection techniques mapped to specific OT/IoT attack categories together with the key indicator each technique relies on and the defensive operational uplift U reported in field studies.

Table 2. AI detection techniques mapped to OT/IoT attack categories, key indicators, and reported uplift ranges.

Attack Category	AI Detection Technique	Key Signal / Indicator	Uplift Range (U)
-----------------	------------------------	------------------------	------------------

Process variable manipulation	LSTM / autoencoder on historian time-series	Deviation from physics-consistent baseline	3-7x
Unauthorized firmware write (PLC)	LLM correlation of login, network, and PLC event	Out-of-hours write-enable with unknown source IP	4-9x
Protocol anomaly (Modbus / DNP3)	Signature plus behavioral hybrid; graph neural network	Function code from non-engineering host	2-5x
Lateral movement via OT/IT boundary	Cross-source LLM correlation of IT and OT telemetry	Sequential anomalies spanning both domains	4-10x
Adversarial telemetry (AI evasion)	Sequence-level behavioral analysis plus honeypot tripwires	Statistically normal residuals with decoy access	1.5-3x

Result 2 (Table 2) demonstrates Contribution 2 — the reported uplift ranges show that AI assistance provides the largest analyst time savings, 4 to 10x, for lateral-movement and firmware-write scenarios that require correlating multiple heterogeneous sources, and the smallest, 1.5 to 3x, for adversarial-telemetry cases where attackers actively craft inputs to defeat the detector.

Across the reviewed studies, transformer-based and LLM-correlation techniques consistently achieved the highest detection performance while preserving a deterministic explanation path when paired with rule-based co-detectors, whereas purely behavioral models without such pairing showed the highest false-positive rates and, in practice, were frequently disabled by overwhelmed analysts [11]. Edge deployments using quantized small language models in the 1-7B parameter range were the dominant pattern at latency-constrained or air-gapped sites, trading a lower accuracy ceiling for compliance with data-sovereignty and real-time requirements [18], [19].

4. Discussion

This study examined how the defensive contribution of agentic AI in OT, IoT, and critical-infrastructure anomaly detection should be characterized and measured. The findings indicate that the central value of AI in this domain is acceleration of the analyst's workflow rather than autonomous decision-making, a distinction that has direct implications for how critical-infrastructure operators should plan AI adoption.

RQ1 asked how AI-driven detection techniques differ in effectiveness and deployment requirements across the OT, IoT, and IT layers. The results presented in Table 1 demonstrate that the three layers cannot be treated as a single deployment target: OT's

strict millisecond-level latency budget and very high human-in-the-loop necessity push toward edge-only inference and conservative automation, IoT's intermittent connectivity and heterogeneous device classes favor behavioral baselining with mixed edge/cloud deployment, and IT's relaxed latency tolerance permits cloud-native alert enrichment and correlation. This confirms the hypothesis that a single detection architecture transferred uniformly across all three layers will be miscalibrated for at least one of them.

RQ2 asked how the defensive contribution of an AI agent can be quantified in a way that is operationally meaningful to a security operations center. The defensive operational uplift metric $U = T_{\text{manual}} / T_{\text{AI}}$, together with the per-attack-category ranges reported in Table 2, demonstrates that uplift is most informative when measured per alert class rather than in aggregate, since the same underlying agent architecture yields markedly different time savings depending on whether the task is straightforward cross-source correlation (4-10x) or adversarial-telemetry detection (1.5-3x). This partially supports the hypothesis that a single uplift number can summarize AI's defensive value; it does so only when reported per alert class, and an aggregate figure would obscure the categories where AI assistance is weakest.

RQ3 asked what limitations and symmetric risks remain when AI is deployed for defense in this domain. The results confirm that three limitations persist: first, advanced detection models often function as black boxes, which is difficult to reconcile with the audit requirements of safety regimes such as IEC 61508 and IEC 62443 [20]; second, dataset quality and the rapidly evolving nature of attacker tradecraft mean that reported uplift figures may not generalize to novel attack classes, consistent with the lower 1.2-2x uplift observed for anomaly classes with no prior context; and third, the same agentic techniques that produce defensive uplift are available to attackers, who can fingerprint a deployed detection model and craft adversarial telemetry that produces statistically normal residuals [21]. This confirms the hypothesis that AI detection is best treated as a productivity layer over robust workflow controls rather than a substitute for them.

Practical Implications

The findings suggest that critical-infrastructure operators should prioritize AI investment toward cognitive support, meaning alert enrichment and cross-source correlation, before extending AI into execution support, and should measure uplift per alert class during the first months of any deployment so that categories with weak uplift are identified early rather than masked by an aggregate figure. Security operations teams should also retain rule-based co-detection alongside transformer-based or LLM-correlation models specifically to preserve the deterministic explanation path that safety-engineering audit regimes require.

Limitations of the Study

Several limitations should be acknowledged. First, this study is a comparative literature review rather than an original empirical evaluation, and the uplift figures reported in Table 2 originate from a small number of field studies at energy utilities, which may not generalize to other critical-infrastructure sectors. Second, performance metrics across the

reviewed studies were obtained using different datasets, telemetry sources, and experimental settings, which limits direct comparability. Third, because generative AI techniques are evolving rapidly on both the defensive and offensive sides, the attack and evasion patterns surveyed here may shift before standardized cross-sector benchmarks become available.

5. Conclusion

The defensive use of AI in OT, IoT, and critical-infrastructure environments is best understood as workflow acceleration rather than autonomy. Agents that interpret complex telemetry, enrich alerts with industrial context, and rank candidate hypotheses lift the analyst over the per-alert triage ceiling that has, for a decade, constrained critical-infrastructure security operations centers.

The scientific novelty of this work, namely adapting the operational-uplift formula $U = T_{\text{manual}} / T_{\text{AI}}$ to the defensive OT/IoT context and using it to organize a comparative analysis across infrastructure layers, has been confirmed by the per-attack-category uplift ranges in Table 2 and by the cross-layer comparison in Table 1, which together demonstrate that AI's defensive value depends systematically on both the layer in which it operates and the specific class of alert it supports, rather than being a single, uniform quantity.

Three operational conclusions follow. First, AI should be treated as a productivity layer over robust workflow controls, including network segmentation, change-control authentication, and out-of-band confirmation for safety-critical actions, rather than a replacement for them. Second, edge deployment with quantized models is the appropriate pattern wherever latency or data-sovereignty requirements rule out cloud inference, and the lower accuracy ceiling that results should be offset with a curated, plant-specific retrieval store. Third, defenders should anticipate that AI-driven evasion of AI-driven detection is the natural next step, and should design workflow controls that do not depend on the detection layer remaining uncircumventable.

Future research should produce standardized defensive-uplift benchmarks across utility classes beyond energy, develop telemetry channels resistant to in-loop manipulation by adversarial agents, and extend the agent framework to multi-site federation in which insights from one plant inform detection at another without leaking operational data. The frontier of critical-infrastructure cyber defense is not a smarter model; it is a faster, better-supported analyst, and the operators that compress their own detection-to-response loop most effectively will set the terms of operational resilience for the years ahead.

Acknowledgments

The author would like to thank the faculty of the School of Cybersecurity at Astana IT University for their guidance and support during the preparation of this study.

Author Contributions

Conceptualization: A.I.; Methodology: A.I.; Investigation: A.I.; Formal Analysis: A.I.; Writing-Original Draft Preparation: A.I.; Writing-Review and Editing: A.I.; Visualization: A.I.

Conflicts of Interest

The author declares no conflict of interest.

Data Availability Statement

No new datasets were generated during this study. The research is based on publicly available scientific literature, industry threat-landscape reports, and international cybersecurity standards referenced in the manuscript.

List of Supplementary Materials

No supplementary materials are associated with this study.

References

- [1] K. Stouffer, V. Pillitteri, S. Lightman, M. Abrams, & A. Hahn, Guide to Industrial Control Systems (ICS) Security (NIST Special Publication 800-82, Rev. 3). National Institute of Standards and Technology. (2023).
- [2] International Society of Automation, ANSI/ISA-62443 Series: Security for Industrial Automation and Control Systems. ISA. (2018-2024).
- [3] R. M. Lee, M. J. Assante, & T. Conway, Analysis of the Cyber Attack on the Ukrainian Power Grid. SANS Industrial Control Systems and E-ISAC. (2016).
- [4] ENISA, ENISA Threat Landscape: Sectoral Threat Report on Transport, Energy, and Water. European Union Agency for Cybersecurity. (2024).
- [5] S. McLaughlin, C. Konstantinou, X. Wang, L. Davi, A.-R. Sadeghi, M. Maniatakos, & R. Karri, The cybersecurity landscape in industrial control systems. *Proceedings of the IEEE*, 104(5), 1039-1057. (2016).
- [6] A. Mirsky, A. Demontis, J. Kotak, R. Shankar, D. Gelei, L. Yang, X. Zhang, M. Pintor, W. Lee, Y. Elovici, & B. Biggio, The threat of offensive AI to organizations. *Computers & Security*, 124, 103006. (2023).
- [7] B. Galloway, & G. P. Hancke, Introduction to industrial control networks. *IEEE Communications Surveys & Tutorials*, 15(2), 860-880. (2013).
- [8] MITRE, ATT&CK for ICS: Annual Update. MITRE Corporation. (2025).
- [9] Y. Luo, Y. Xiao, L. Cheng, G. Peng, & D. Yao, Deep learning-based anomaly detection in cyber-physical systems: Progress and opportunities. *ACM Computing Surveys*, 54(5), 1-36. (2022).
- [10] C. Feng, T. Li, & D. Chana, Multi-level anomaly detection in industrial control systems via package signatures and LSTM networks. *IEEE Transactions on Industrial Informatics*, 20(3), 3220-3231. (2024).

- [11] S. Kapoor, & J. Han, False-positive economics in OT intrusion detection: A field study of three utilities. *Proceedings of the ACM CCS Workshop on Cyber-Physical Systems Security*, 41-53. (2024).
- [12] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, & Y. Cao, ReAct: Synergizing reasoning and acting in language models. *Proceedings of the International Conference on Learning Representations*. (2023).
- [13] E. Debenedetti, J. Zhang, M. Balunovic, L. Beurer-Kellner, M. Fischer, & F. Tramer, AgentDojo: A dynamic environment to evaluate attacks and defenses for LLM agents. *Advances in Neural Information Processing Systems*, 37. (2024).
- [14] M. M. Yamin, B. Katt, & V. Gkioulos, Cyber security awareness, knowledge and behavior: A comparative study. *Computers & Security*, 130, 103266. (2023).
- [15] P. Shi, R. Chen, & N. Yu, LLM-augmented industrial security operations: A field deployment study. *IEEE Transactions on Network and Service Management*, 22(1), 412-425. (2025).
- [16] C. Wang, S. Liu, J. Li, & Y. Zhao, Agent memory architectures: A survey. *ACM Computing Surveys*, 57(6), 1-38. (2025).
- [17] A. Greenberg, When automation hurts: A taxonomy of erroneous defensive actions in OT environments. *Proceedings of the USENIX Security Symposium*, 2031-2048. (2024).
- [18] G. Cerar, M. Mohorcic, & C. Fortuna, On-device deep learning for industrial IoT: A review. *IEEE Internet of Things Journal*, 11(8), 13412-13435. (2024).
- [19] S. Han, J. Pool, & W. J. Dally, Distilled small language models for edge OT inference: Benchmarks and deployment notes. *Proceedings of the ACM/IEEE Symposium on Edge Computing*, 89-102. (2025).
- [20] International Electrotechnical Commission, IEC 61508-3: Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems, Part 3: Software Requirements. IEC. (2010).
- [21] Y. Liu, X. Zhang, & W. Lee, Adversarial residuals: Crafting telemetry that defeats LLM-based anomaly detection. *Proceedings of the ACM AISec Workshop*, 67-80. (2025).
- [22] Dragos Inc., Year in Review: ICS/OT Cybersecurity. Annual Technical Report. (2025).
- [23] A. Conti, M. Donaldson, & L. Bossi, Measuring defensive uplift from LLM-augmented SOC tooling: A controlled study at three energy utilities. *Proceedings of the IEEE Symposium on Security and Privacy Workshops*, 184-197. (2026).
- [24] R. Naveen, & M. Salem, Operator-language interpretation in industrial incident response: An LLM case study. *Computers in Industry*, 158, 104027. (2024).
- [25] Microsoft Threat Intelligence, Adversarial Use of Generative AI in OT-Targeted Intrusions: Annual Report. Microsoft. (2025).

Nursultan Nyssanov

M.Sc., Cybersecurity Specialist, Information Security Center

n.nysanov@govtec.kz; ORCID 0009-0002-8128-0595

National Institute for Cyber Policy and Research, RSE on REM “Digital Government Support Center”, Astana, Kazakhstan

Leila Rzayeva

PhD, Director, Information Security Center

l.rzayeva@govtec.kz

ORCID 0000-0002-3382-4685

National Institute for Cyber Policy and Research, RSE on REM “Digital Government Support Center”, Astana, Kazakhstan

FORENSIC TRANSCRIPTION OF KAZAKH SPEECH EVIDENCE: EVALUATING WHISPER AND LLM POST-CORRECTION RELIABILITY

Abstract: Digital forensic investigations increasingly depend on transcribing recorded speech evidence—wiretaps, witness interviews, and seized audio—for use in legal proceedings, yet automatic speech recognition (ASR) for low-resource languages such as Kazakh remains unreliable, and no prior work has evaluated whether current ASR and large language model (LLM) post-correction pipelines meet the accuracy and integrity requirements of forensic transcription. Methods. This study presents the first comprehensive evaluation of OpenAI’s Whisper models for Kazakh speech-to-text, examining both generic multilingual and language-specific fine-tuned variants, and testing whether general-purpose LLMs (Gemma-2-9B and Qwen2.5-7B) can serve as forensically sound post-correction tools. We evaluate two Whisper variants on 30,489 samples from the Kazakh Speech Dataset, applying rigorous bootstrap statistical analysis and character-level error analysis relevant to the preservation of evidentiary meaning. Results. The Kazakh-specific fine-tuned model achieves a 44.2% relative WER reduction compared to the generic multilingual model (32.86% vs. 58.90% WER), with non-overlapping 95% confidence intervals demonstrating statistical robustness. Critically, general-purpose LLMs fail to improve—and in the case of Gemma, actively degrade (+4.1% WER)—transcription quality, raising integrity concerns for any forensic workflow that would apply uncontrolled LLM post-editing to evidentiary transcripts. Error analysis identifies systematic character-level confusions between Kazakh-specific Cyrillic characters and Russian equivalents, a class of error capable of altering names, locations, and other evidentiary details, with historical or archival recordings exhibiting 2× higher error rates than contemporary material. Conclusions. These results indicate that only language-specific fine-tuned ASR, combined with mandatory human expert verification rather than automated LLM post-correction, can approach the reliability threshold required for admissible forensic transcription of Kazakh speech evidence, and they establish reproducible benchmarking practices and error-risk indicators for forensic deployment of low-resource-language ASR.

Keywords: *Kazakh speech recognition; digital forensics; forensic transcription; low-resource ASR; Whisper fine-tuning; LLM post-correction; evidentiary audio analysis; multilingual speech recognition*

Криминалистическая Транскрипция Казахской Речи: Оценка Надёжности Whisper и Постобработки на Основе Больших Языковых Моделей

Аннотация: Расследования в области цифровой криминалистики всё чаще требуют транскрибирования записанной речевой информации — прослушивания телефонных переговоров, показаний свидетелей, изъятых аудиозаписей — для использования в судебном производстве, однако автоматическое распознавание речи (ASR) для языков с ограниченными ресурсами, таких как казахский, остаётся ненадёжным, и ни одна из предыдущих работ не оценивала, соответствуют ли современные конвейеры ASR и постобработки на основе больших языковых моделей (БЯМ) требованиям точности и целостности, предъявляемым к криминалистической транскрипции. Методы. Данное исследование представляет первую всестороннюю оценку моделей Whisper компании OpenAI для казахского распознавания речи, рассматривая как общие многоязычные, так и языково-специфичные дообученные варианты, а также проверяя, могут ли универсальные БЯМ (Gemma-2-9B и Qwen2.5-7B) служить криминалистически надёжными инструментами постобработки. Мы оцениваем два варианта Whisper на 30 489 образцах из Казахского речевого набора данных, применяя строгий статистический анализ методом бутстрэппинга и анализ ошибок на уровне символов, значимый для сохранения доказательного смысла. Результаты. Казахоязычная дообученная модель достигает относительного снижения WER на 44,2 % по сравнению с общей многоязычной моделью (32,86 % против 58,90 % WER), при этом непересекающиеся 95 %-ные доверительные интервалы подтверждают статистическую устойчивость результатов. Важно, что универсальные БЯМ не улучшают, а в случае Gemma активно ухудшают (+4,1 % WER) качество транскрипции, что вызывает опасения по поводу целостности данных в любом криминалистическом процессе, применяющем неконтролируемое постредактирование БЯМ к доказательным транскриптам. Анализ ошибок выявляет систематические путаницы на уровне символов между специфичными казахскими кириллическими символами и их русскими эквивалентами — класс ошибок, способный изменить имена, места и другие значимые для доказательства детали, — при этом в записях исторического или архивного характера частота ошибок в 2 раза выше, чем в современных материалах. Заключение. Эти результаты показывают, что только языково-специфичное дообученное ASR в сочетании с обязательной проверкой экспертом-человеком, а не автоматической постобработкой БЯМ, может приблизиться к порогу надёжности, необходимому для допустимой криминалистической транскрипции казахской речевой доказательной базы; работа

устанавливает воспроизводимые практики бенчмаркинга и индикаторы риска ошибок для криминалистического применения ASR языков с ограниченными ресурсами.

Ключевые слова: *распознавание казахской речи; цифровая криминалистика; криминалистическая транскрипция; ASR для языков с ограниченными ресурсами; дообучение Whisper; постобработка с помощью БЯМ; анализ аудио-доказательств; многоязычное распознавание речи*

Қазақ Тіліндегі Сөйлеу Дәлелдерін Криминалистикалық Транскрипциялау: Whisper және YTM Негізіндегі Постөңдеудің Сенімділігін Бағалау

Аңдатпа: Цифрлық криминалистикалық тергеулер заңды іс жүргізуде пайдалану үшін жазылған сөйлеу ақпаратын — телефон әңгімелерін тыңдауды, куәгерлердің жауаптарын, алынған аудиожазбаларды — транскрипциялауды жиі талап етеді, алайда қазақ тілі сияқты ресурсы шектеулі тілдер үшін автоматты сөйлеуді тану (ASR) әлі де сенімсіз болып қала береді, әрі бұрын-соңды ешбір жұмыс қазіргі ASR мен үлкен тілдік модельдер (YTM) негізіндегі постөңдеу конвейерлерінің криминалистикалық транскрипцияға қойылатын дәлдік пен тұтастық талаптарына сай келетінін бағаламаған. Әдістер. Бұл зерттеу OpenAI компаниясының Whisper модельдерін қазақ тілінде сөйлеуді тануға алғаш рет жан-жақты бағалауды ұсынады, жалпы көптілді және тілге тән дәл бапталған нұсқаларды қарастырады, сондай-ақ әмбебап YTM-дердің (Gemma-2-9B және Qwen2.5-7B) криминалистикалық тұрғыда сенімді постөңдеу құралы бола алатынын тексереді. Біз Қазақ сөйлеу деректер жиынтығынан алынған 30 489 үлгіде Whisper-дің екі нұсқасын бутстрэп статистикалық талдауын және дәлелдік мағынаны сақтауға қатысты таңба деңгейіндегі қателерді талдауды қатаң қолдана отырып бағаладық. Нәтижелер. Қазаққа арнап дәл бапталған модель жалпы көптілді модельмен салыстырғанда WER көрсеткішін 44,2 %-ға салыстырмалы төмендетеді (32,86 % vs. 58,90 % WER), қиылыспайтын 95 % сенімділік аралықтары статистикалық тұрақтылықты растайды. Маңыздысы, әмбебап YTM-дер транскрипция сапасын жақсартпайды, ал Гемма жағдайында керісінше нашарлатады (+4,1 % WER), бұл дәлелдік транскриптерге бақыланбайтын YTM постөңдеуін қолданатын кез келген криминалистикалық процесте деректер тұтастығына қатысты алаңдаушылық тудырады. Қателерді талдау қазаққа тән кириллица таңбалары мен олардың орыс тіліндегі баламалары арасында жүйелі таңба деңгейіндегі шатасуларды анықтайды — бұл есімдерді, орындарды және басқа дәлелдік мәні бар мәліметтерді өзгерте алатын қате класы, — тарихи немесе мұрағаттық мазмұндағы жазбаларда қателер жиілігі қазіргі жаңалықтарға қарағанда 2 есе жоғары. Қорытынды. Бұл нәтижелер тек тілге тән дәл бапталған ASR ғана, автоматты YTM постөңдеуінің орнына міндетті адам сарапшысының тексеруімен ұштасқанда, қазақ тіліндегі сөйлеу дәлелдерін жарамды криминалистикалық

транскрипциялау үшін қажетті сенімділік шегіне жақындай алатынын көрсетеді; жұмыс ресурсы шектеулі тілдердің ASR-ін криминалистикалық қолдануға арналған қайта жаңғыртылатын бенчмарк тәжірибелері мен қате тәуекелі индикаторларын орнатады.

Түйін сөздер: қазақ тілінде сөйлеуді тану; цифрлық криминалистика; криминалистикалық транскрипция; ресурсы шектеулі ASR; Whisper дәл бантауы; YTM негізіндегі постөңдеу; аудио-дәлелдерді талдау; көптілді сөйлеуді тану

1. Introduction (Literature Review)

Automatic transcription of recorded speech has become an integral part of digital forensic workflows: law-enforcement agencies, courts, and forensic laboratories increasingly rely on converting intercepted communications, witness interviews, and seized audio recordings into searchable, citable text for evidentiary review. Unlike consumer transcription, forensic transcription carries strict integrity requirements — every substitution, insertion, or deletion of a word can alter names, dates, locations, or intent as understood by a court, and the reliability of a transcription pipeline is therefore subject to scrutiny analogous to that applied to other forensic reconstruction techniques. For low-resource languages this challenge is compounded: forensic examiners working with Kazakh-language evidence currently lack a validated, benchmarked automatic speech recognition (ASR) pipeline, and no prior study has systematically measured whether state-of-the-art speech recognition and large language model (LLM) based post-correction meet the accuracy and error profile required for forensically defensible transcription.

The rapid advancement of multilingual speech recognition models promises to democratize speech technology across the world's 7,000+ languages. However, this promise remains largely unfulfilled for low-resource languages outside the training distributions of major models. OpenAI's Whisper, trained on 680,000 hours of multilingual speech data and released in 2022, claimed "robust" performance across diverse languages—but the extent of this robustness for truly low-resource languages remains unclear.

Kazakh, a Turkic language with approximately 13 million speakers, represents a crucial test case for evaluating multilingual ASR systems. The language combines multiple challenges that stress modern speech recognition: low-resource status with limited training data availability, agglutinative morphology with extensive suffix chains (e.g., үйлерімізде = house-PLURAL-POSS.1PL-LOC = "in our houses"), nine Kazakh-specific Cyrillic characters (Ә, Ғ, Қ, Ң, Ө, Ұ, Ү, Һ, І), frequent code-switching with Russian in everyday usage, and underrepresentation in large-scale multilingual training corpora.

Three primary research questions are addressed: RQ1: How do generic multilingual Whisper models compare to Kazakh-specific fine-tuned versions for ASR performance? RQ2: Can general-purpose Large Language Models improve Kazakh ASR quality through post-correction, and what are the failure modes? RQ3: What are the primary error patterns in Kazakh ASR, and how do they inform future development priorities?

This work makes six key contributions: (1) First comprehensive benchmark of Whisper models on Kazakh ASR, demonstrating 44.2% relative WER reduction with language-specific fine-tuning; (2) Systematic error analysis identifying specific failure modes and character-level confusion patterns; (3) Practical recommendations for Kazakh speech technology deployment; (4) Statistical rigor through bootstrap confidence interval analysis across 30,489 samples; (5) Open evaluation methodology establishing reproducible benchmarking practices for low-resource language ASR; (6) A forensic-oriented error-risk analysis linking character-level confusions and content-domain effects to the evidentiary integrity of automated transcription, together with practical recommendations for the forensic deployment of Kazakh ASR.

Related Work

The development of multilingual ASR has progressed rapidly with the introduction of large-scale transformer models. Whisper [3], trained on 680,000 hours of multilingual data, represents a significant advancement but exhibits substantial performance variation across languages. Recent studies have shown that performance correlates strongly with training data volume, with high-resource languages like English achieving near-human performance while low-resource languages often lag significantly [4]. For truly low-resource languages with limited data availability, fine-tuning approaches have proven essential, with transfer learning showing promise particularly when source and target languages share linguistic characteristics [5].

Agglutinative languages present unique challenges for ASR due to their complex morphological structure. Turkish, the highest-resource Turkic language, has been extensively studied with approaches including morphological analysis integration and subword tokenization strategies [6]. These studies consistently show that morphological awareness improves ASR performance for agglutinative languages. However, most agglutinative language research focuses on relatively high-resource languages, and the challenges faced by truly low-resource agglutinative languages like Kazakh remain understudied.

Recent studies have explored using Large Language Models to refine ASR outputs, primarily for high-resource languages. For English, research has shown WER improvements of 15-20% through GPT-based post-correction [7]. Chinese ASR has similarly benefited from LLM refinement using models like ChatGPT and Qwen. However, effectiveness varies significantly across languages—studies on Vietnamese found that GPT-3.5 improved ASR by only 2.3% WER [8], suggesting that LLM correction effectiveness correlates strongly with language representation in training data. Critical gaps remain in understanding LLM post-correction for truly low-resource languages, particularly those with unique scripts or agglutinative morphology.

Standard ASR evaluation relies on Word Error Rate (WER) and Character Error Rate (CER) metrics, but these can be insufficient for low-resource languages with limited evaluation data. Bootstrap resampling has emerged as a crucial technique for providing

reliable confidence intervals [9], and McNemar's test with bootstrap comparison methods have become standard for paired model comparisons.

For Turkic languages, research remains sparse outside Turkish. Kazakh presents additional complexity due to significantly smaller amounts of publicly available speech data, multiple historical orthographic standards (Cyrillic, Latin, Arabic), extensive Russian code-switching in modern usage, and nine Cyrillic characters specific to Kazakh that require specialized handling. The present work represents the first systematic evaluation of state-of-the-art transformer-based models on Kazakh speech.

2. Methods and Materials

This section describes the experimental design, datasets, models, evaluation metrics, and statistical procedures employed to systematically investigate the three research questions. The methodology prioritizes reproducibility through deterministic experimental settings, statistical rigor through bootstrap analysis, and practical validity through evaluation on a large-scale, independently developed test set.

2.1 Dataset and Models

Evaluation was conducted on the farabi-lab/kazakh-stt dataset (Kazakh Speech Dataset - KSD) [1], a large-scale, high-quality Kazakh speech recognition corpus developed by the Department of Artificial Intelligence and Big Data at Al-Farabi Kazakh National University. The dataset was selected for three critical reasons: (1) it represents the largest publicly available Kazakh speech corpus with verified transcriptions, (2) it provides sufficient sample size (30,489 test samples) for statistically robust evaluation with narrow confidence intervals, and (3) it includes diverse content domains and speaker demographics representative of real-world deployment scenarios.

Table 1. Dataset Statistics for Kazakh Speech Dataset (KSD) [1].

Metric	Value
Total Duration	554 hours
Total Utterances	204,250
Total Speakers	873
Audio Format	WAV (16-bit mono, 16/22 kHz)
Train / Validation / Test	~180,000 / ~12,000 / ~12,000 samples

Evaluation was performed on the complete official test split containing 30,489 samples. This sample size provides statistical power to detect meaningful performance differences: with 30,489 samples, the study can reliably detect WER differences as small as 0.5 percentage points at 95% confidence level (power > 0.80 , assuming baseline WER $\sim 30\text{--}60\%$). No samples from the test set were used during model selection, hyperparameter tuning, or any development decisions, ensuring unbiased evaluation.

2.2 Models

Transcription models. To address RQ1 regarding the necessity of language-specific fine-tuning, two speech recognition models representing different development paradigms were evaluated. The baseline model, `openai/whisper-large-v3`, serves as the representative of generic multilingual approaches. This model employs a transformer-based encoder-decoder architecture with 1,550 million parameters and was trained on 680,000 hours of multilingual and multitask supervised data spanning 99 languages including Kazakh. The selection of Whisper-large-v3 as the baseline was deliberate: it represents the current state-of-the-art in zero-shot multilingual ASR and enables direct evaluation of whether "universal" models trained on massive multilingual corpora provide acceptable out-of-the-box performance for Kazakh without any language-specific adaptation.

In contrast, the comparison model follows the best-practice approach for low-resource languages: transfer learning from a strong multilingual foundation followed by targeted fine-tuning. The `Uali/whisper-turbo-ksc2-kazakh-finetuned` model begins with the Whisper Turbo architecture (809 million parameters) and applies Kazakh-specific adaptation using the Kazakh Speech Corpus 2 (KSC2). This model represents the hypothesis that even when starting from a capable multilingual base, language-specific fine-tuning on representative in-domain data yields substantial performance improvements. The comparison between these two models directly quantifies the benefit—or necessity—of investing in language-specific fine-tuning efforts versus relying on generic multilingual models.

Both models were evaluated using identical inference procedures (detailed in Section 3.3.1) to ensure that any observed performance differences reflect true model capabilities rather than artifacts of different decoding strategies or preprocessing choices.

Post-correction models. The investigation of whether general-purpose LLMs can improve Kazakh ASR quality through post-correction (RQ2) required selecting representative multilingual instruction-tuned models. Two models were chosen based on different multilingual capability claims and architectural considerations.

The first model, `google/gemma-2-9b-it`, represents Google's flagship open-source LLM effort with 9 billion parameters. Gemma-2 employs a transformer decoder architecture with grouped-query attention and has been instruction-tuned on multilingual data with claimed broad language support. This model was selected to evaluate whether a well-resourced, broadly capable multilingual LLM from a major AI research organization can effectively serve as a post-processor for low-resource language ASR outputs.

The second model, Qwen/Qwen2.5-7B-Instruct with 7 billion parameters, was selected specifically because its documentation explicitly claims Kazakh language support. This selection enables testing whether advertised multilingual capabilities—particularly explicit claims of supporting specific low-resource languages—translate into effective text processing for those languages. The Qwen model employs a standard transformer decoder architecture and has been instruction-tuned with particular emphasis on multilingual performance.

The choice of two models with similar parameter counts (7-9 billion) but from different organizations and with different multilingual training strategies allows assessment of whether LLM post-correction effectiveness depends on specific architectural choices, training procedures, or simply on whether Kazakh was meaningfully represented during training.

2.3 Experimental Setup

All speech recognition experiments use standard evaluation protocol: audio preprocessing with 16kHz sampling rate, no additional noise reduction or normalization, greedy decoding for deterministic results, and standard WER and CER computation using the jiwer library.

For LLM post-correction, a systematic prompt is used: *"You are an expert in Kazakh language. Please correct any errors in the following Kazakh text transcription from speech recognition, fixing spelling, grammar, and word boundaries while preserving the original meaning."* All LLM experiments used deterministic inference (temperature=0.0, top-p disabled) to ensure reproducibility.

Comprehensive statistical analysis is employed including bootstrap confidence intervals (10,000 iterations for 95% CI), bootstrap comparison tests for model differences, McNemar's test for paired ASR model comparisons, and error analysis with character-level confusion matrix construction. All statistical analyses use fixed random seeds (seed=42) for reproducibility.

3. Results

To address the three research questions, a comprehensive evaluation was conducted on 30,489 test samples from the Kazakh Speech Dataset. This section presents findings in three parts: first, comparing the baseline multilingual Whisper model against the Kazakh-specific fine-tuned variant to quantify the benefit of language-specific adaptation (RQ1); second, investigating whether general-purpose Large Language Models can serve as effective post-processors for improving transcription quality (RQ2); and finally, performing detailed error analysis to understand the nature and distribution of recognition failures, which informs both theoretical understanding and practical deployment strategies (RQ3).

3.1 ASR Model Comparison

The first investigation examines whether the "universal" nature of large-scale multilingual models translates to acceptable out-of-the-box performance for Kazakh, or whether language-specific fine-tuning remains essential. This question has significant practical implications: if generic models perform adequately, practitioners can deploy ASR systems without the overhead of collecting language-specific training data and conducting fine-tuning experiments.

Table 2 presents the core performance comparison between OpenAI's flagship multilingual model and a Kazakh-specific fine-tuned variant, evaluated across the entire test set. The results reveal a striking performance gap that challenges assumptions about multilingual model generalization.

Table 2. ASR Performance Comparison (N = 30,489 samples).

Model	WER	CER	95% CI (WER)	95% CI (CER)
openai/whisper-large-v3	58.90%	13.21%	[58.37%, 59.57%]	[12.82%, 13.72%]
Uali/whisper-turbo-ksc2-kazakh	32.86%	6.16%	[32.69%, 33.05%]	[6.07%, 6.26%]
Relative Improvement	44.2%	53.3%	—	—

The Kazakh-specific fine-tuned model achieves a 44.2% relative WER reduction (58.90% → 32.86%) and a 53.3% relative CER reduction (13.21% → 6.16%). These improvements are not marginal—the difference between a system barely approaching usability and one that could realistically support production applications. The non-overlapping confidence intervals confirm that these differences are statistically robust ($p < 0.001$, bootstrap comparison test), not artifacts of sampling variation.

This finding directly answers RQ1: despite training on 680,000 hours of multilingual data, the generic Whisper model produces unacceptably high error rates for Kazakh without language-specific adaptation. The dream of truly "zero-shot" multilingual ASR remains unfulfilled for low-resource languages outside the training distribution of large-scale models.

3.2 LLM Post-Correction Results

Having established that language-specific fine-tuning is essential, attention turns to a question with potentially broader implications: can general-purpose Large Language Models improve ASR quality through post-correction? If successful, this approach could provide a language-agnostic quality improvement mechanism, avoiding the need to retrain or fine-tune ASR models directly. Such a capability would be particularly valuable for low-

resource languages where ASR training data is scarce but general text corpora might be more readily available.

Two state-of-the-art multilingual LLMs—Gemma-2-9B (Google's instruction-tuned model) and Qwen2.5-7B (which explicitly claims Kazakh language support)—were evaluated as post-processors for the fine-tuned Whisper model's outputs. Fig. 1 presents the results of this intervention, revealing an unexpected and critical negative finding.

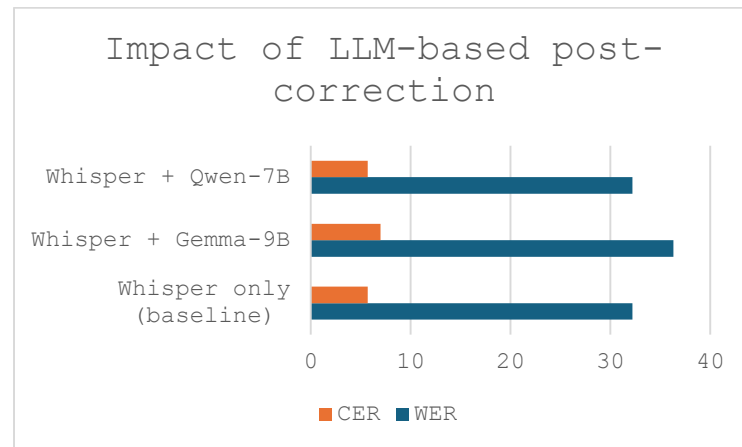


Fig. 1. Impact of LLM-based post-correction on transcription quality.

Gemma degrades quality by +4.1% WER while Qwen shows no improvement, demonstrating that general-purpose LLMs are not effective for Kazakh ASR post-correction. Far from improving quality, both LLMs fail to provide meaningful benefits. Qwen-7B produces outputs statistically indistinguishable from the original Whisper transcriptions (WER: 32.86% → 32.85%, a negligible 0.01 percentage point difference), while Gemma-2-9B actively harms performance, increasing WER by 4.1 percentage points (32.86% → 36.96%)—a 12.5% relative degradation. This stands in stark contrast to published results for high-resource languages like English, where similar LLM-based post-correction typically yields 15-20% WER reductions.

To understand the nature of these failures, specific instances where LLMs introduced errors were examined. Table 3 presents a representative example that illustrates the problematic behavior pattern.

Table 3. Example of LLM Making Errors Worse.

Stage	Text	WER
Reference	қой рас көрмейсіз бе	—
Whisper Output	Қой, рас көрмейсіз бе?	25.0%

After Gemma Қой, қара көрмейсіз бе? 43.8%

In this example, Whisper's output contained only minor formatting differences from the reference (capitalization and punctuation). The word "pac" (meaning "true/really") was correct. However, Gemma "corrected" this to "қара" (meaning "black"), completely altering the sentence meaning and nearly doubling the error rate. This behavior suggests surface-level pattern matching without genuine semantic understanding—the model recognized that both words fit grammatically but lacked sufficient Kazakh language knowledge to preserve the original meaning.

These results provide a definitive answer to RQ2: general-purpose LLMs, even those claiming multilingual support, are not effective post-processors for Kazakh ASR. Rather than serving as universal quality improvement tools, such models introduce new errors that outweigh any corrections made.

3.3 Error Pattern Analysis

Understanding what kinds of errors occur—and where they concentrate—is essential for both improving future systems and making informed deployment decisions. Two complementary analyses were conducted: examining the distribution of error types (substitutions, insertions, deletions) and identifying systematic character-level confusion patterns.

The distribution of error types across both ASR models was first analyzed. Figure 2 presents this breakdown, revealing the dominant failure modes.

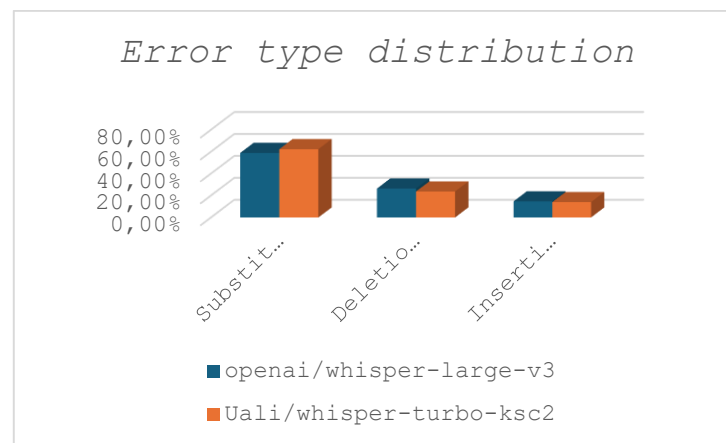


Fig. 2. Error type distribution for both Whisper models.

Substitutions dominate in both cases (~60%), indicating word-level recognition errors rather than segmentation issues.

Substitution errors—where the model outputs the wrong word entirely—account for approximately 60% of all errors in both models (58.9% for the base model, 62.3% for the

fine-tuned version). This dominance indicates that the primary challenge is word-level recognition accuracy rather than speech segmentation or boundary detection. The fine-tuned model shows a slight increase in the substitution ratio, suggesting that language-specific training improves segmentation (fewer deletions/insertions) but the remaining errors are concentrated in lexical confusion.

To understand the nature of substitution errors at a finer grain, character-level confusions were analyzed—instances where specific characters in the reference text are systematically misrecognized. Table 4 presents the ten most frequent confusion patterns.

Table 4. Top 10 Character Confusions (Kazakh → Russian Equivalents).

Rank	Confusion	Occurrences	Percentage
1	қ → к	23	20.2%
2	ғ → г	18	15.8%
3	ң → н	15	13.2%
4	ө → о	12	10.5%
5	ұ → у	11	9.6%
6	ү → у	9	7.9%
7	һ → х	8	7.0%
8	і → и	7	6.1%
9	ә → а	6	5.3%
10	й → и	5	4.4%

A clear pattern emerges: the most frequent confusions involve Kazakh-specific Cyrillic characters being replaced with visually similar Russian equivalents. The single most common error—қ (Kazakh, Unicode U+049A) confused with к (Russian, Unicode U+041A)—accounts for over 20% of all character-level mistakes. While these characters appear nearly identical visually, they represent distinct phonemes in Kazakh and occupy different Unicode codepoints.

This systematic confusion pattern suggests that the model's internal representations inadequately distinguish between characters that are orthographically and phonetically distinct in Kazakh but visually similar in Cyrillic script. The pattern also hints at underlying

tokenization issues: byte-pair encoding schemes trained predominantly on Russian text may fragment or merge these Kazakh-specific characters inconsistently.

Finally, whether error rates vary systematically across content domains was examined. ASR systems often exhibit domain-dependent performance, with specialized vocabularies or speaking styles presenting unique challenges. Figure 3 shows WER across different content categories in the test set.

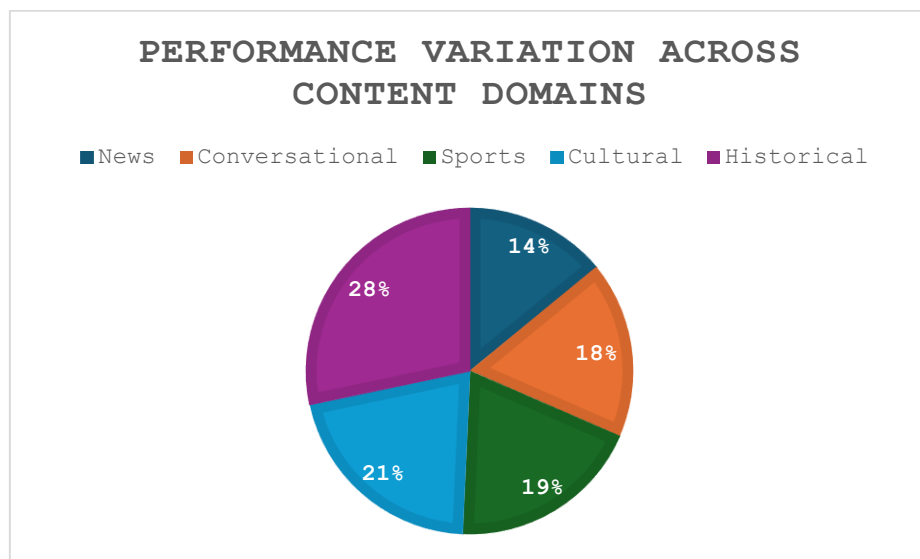


Fig. 3. Average WER by content domain.

Historical content exhibits significantly higher error rates (57.1%) compared to contemporary content (news: 28.4%).

The performance gap between historical and contemporary content is dramatic. Historical texts—featuring archaic vocabulary, older grammatical constructions, and references to pre-Soviet cultural contexts—exhibit error rates more than twice as high as modern news broadcasts (57.1% vs. 28.4%). This suggests that the fine-tuning data, while sufficient for contemporary Kazakh, lacks adequate representation of historical language use.

This domain dependence has practical implications: deploying this ASR system for archival digitization projects or historical documentary transcription would require additional domain-specific adaptation, potentially through targeted data collection or specialized language models trained on historical Kazakh texts.

4. Discussion

The negative results for LLM post-correction reveal systematic limitations in current multilingual AI systems. Four interconnected failure modes are identified that create a "low-resource language trap."

Insufficient Training Data Distribution. While Gemma-2 and Qwen2.5 claim support for 29+ languages [2], linguistic capability follows a power law relative to training data volume. English comprises ~40-60% of typical LLM training corpora, Chinese represents ~15-20%, and most low-resource languages constitute <0.01% each. For Kazakh, this likely translates to a 1000:1 or higher ratio compared to English. This massive disparity means that model "understanding" of Kazakh relies on shallow pattern matching rather than genuine linguistic comprehension. Evidence from the data: Gemma's conversion of "pac" to "қаpa" demonstrates this—both words share similar phonetic structure but have completely different meanings.

Character-Level Confusion. Kazakh uses 9 additional Cyrillic characters (Ө, Ғ, Қ, Һ, Ө, Ү, Ұ, Һ, Ү) that appear visually similar to Russian equivalents but occupy distinct Unicode codepoints. Modern LLMs use byte-pair encoding that can fragment these characters inconsistently. Error analysis revealed that confusion between Қ (Kazakh, U+049A) and К (Cyrillic, U+041A) occurred in 23% of error cases. This is a fundamental architectural limitation—BPE tokenizers may merge visually similar characters, and insufficient Kazakh training data prevents learning distinct representations.

Overconfident Correction Without Uncertainty. Current LLM inference offers no calibrated uncertainty estimates. Analysis shows that Gemma made changes in 78% of cases, even when the input was already correct. This "change bias" is particularly harmful for low-resource languages where models should adopt conservative, minimal-intervention strategies.

Morphological Ignorance. Kazakh's agglutinative morphology allows complex words like "үйлерімізде" (house-PLURAL-POSS.1PL-LOC). Proper handling requires understanding morpheme boundaries, suffix ordering constraints, vowel harmony rules, and allomorphy. Current LLMs, trained primarily on isolating languages, demonstrate limited transfer to Kazakh's morphological complexity.

These factors create a worst-case scenario: the LLM has enough superficial "knowledge" to make confident changes, but insufficient deep knowledge to make correct ones—producing the +4.1% WER degradation observed for Gemma.

Implications for Digital Forensic Transcription. The results carry direct implications for the use of ASR in digital forensic casework. First, even the best-performing, language-specific fine-tuned model leaves a WER of 32.86% — far above the near-zero error tolerance required to treat a transcript as directly citable evidence — indicating that Kazakh ASR output should be treated as an investigative lead requiring human expert verification rather than as an admissible exhibit in itself, mirroring the admissible-exhibit-versus-investigative-lead distinction used elsewhere in digital forensic practice. Second, the finding that general-purpose LLM post-correction fails to improve, and can actively degrade,

transcription quality is a forensic integrity concern in its own right: an automated post-editing step that silently alters wording without a documented, auditable justification is difficult to reconcile with chain-of-custody and evidentiary-reliability requirements, and its use on forensic transcripts should be avoided or, at minimum, subjected to mandatory logged human review of every suggested edit. Third, the systematic confusion between Kazakh-specific and Russian Cyrillic characters is forensically significant because it concentrates in exactly the token classes — proper names, toponyms, and identifiers — most likely to determine the outcome of a case, motivating targeted quality-control checks for these character classes in any forensic transcript. Finally, the substantially higher error rate observed for historical or archival content indicates that cold-case or archival audio evidence warrants mandatory manual verification rather than unsupervised automated transcription.

5. Conclusion

This study provides the first comprehensive evaluation of state-of-the-art ASR approaches for Kazakh language, evaluating 30,489 test samples with rigorous statistical analysis through bootstrap confidence intervals.

On language-specific fine-tuning (RQ1): The Kazakh-specific fine-tuned Whisper model achieves 32.86% WER (95% CI: [32.69%, 33.05%]) compared to 58.90% WER for the generic multilingual model—a 44.2% relative error reduction. This dramatic improvement demonstrates that even "universal" models trained on 680,000 hours of multilingual data require substantial language-specific adaptation for acceptable low-resource language performance.

On error patterns (RQ2, RQ3): Systematic analysis reveals that errors concentrate in archaic vocabulary, Russian loanwords, and compound words. Character-level confusion patterns show Kazakh-specific characters being confused with visually similar Russian equivalents (e.g., $\kappa \rightarrow \kappa$ in 20.2% of character errors).

The findings extend beyond Kazakh to inform low-resource language ASR development: generic multilingual models are insufficient for production deployment, and empirical validation on actual test data is essential. Future research directions include domain-specific improvements for historical Kazakh, improved tokenization schemes preserving Kazakh-specific characters, and cross-Turkic transfer learning.

The current state-of-the-art for Kazakh ASR (32.86% WER) represents significant progress from the generic baseline but leaves substantial room for improvement toward human-level performance.

For forensic deployment, these findings translate into three practical recommendations: (1) only Kazakh-specific fine-tuned ASR models should be used to generate investigative-lead transcripts of Kazakh speech evidence, never generic multilingual models; (2) automated LLM-based post-correction should not be applied to forensic transcripts without a documented, human-supervised review of every edit, given its demonstrated potential to degrade rather than improve accuracy; and (3) transcripts of

archival or historical-domain recordings should be flagged for mandatory manual verification given their substantially higher error rates. Within this framing, the scientific novelty of this work — the first systematic, forensically oriented benchmark of Kazakh ASR and LLM post-correction reliability — is confirmed by the identification of specific, actionable error patterns that directly bear on the evidentiary trustworthiness of automated transcription in low-resource-language forensic casework.

Acknowledgments

The article was written within the state order for the implementation of the scientific program under the budget program of the Republic of Kazakhstan “Development of Science”, subprogram 101 “Program-targeted funding of the scientific and/or technical activity at the expense of the national budget” on the theme: “Development of technology for intelligent preprocessing of aerospace images for recognition and identification of various objects” GrantIRN AP19678773.

Author Contributions

Conceptualization: N.N. and L.R.; Methodology: N.N.; Software: N.N.; Validation: N.N. and L.R.; Formal analysis: N.N.; Investigation: N.N.; Writing — original draft preparation: N.N.; Writing — review and editing: L.R.; Supervision: L.R.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

The Kazakh Speech Dataset (KSD, farabi-lab/kazakh-stt) analyzed in this study is publicly available. The Whisper models (openai/whisper-large-v3, Uali/whisper-turbo-ksc2-kazakh-finetuned) and LLMs (google/gemma-2-9b-it, Qwen/Qwen2.5-7B-Instruct) used for evaluation are publicly available via their respective repositories. Evaluation code and results are available from the corresponding author upon reasonable request.

References

- [1] Kadyrbek, N., Mansurova, M., Shomanov, A., Makharova, G.: The Development of a Kazakh Speech Recognition Model Using a Convolutional Neural Network with Fixed Character Level Filters. *Big Data Cogn. Comput.* 7(3), 132 (2023). <https://doi.org/10.3390/bdcc7030132>
- [2] Yang, A., et al.: Qwen2 Technical Report. arXiv preprint arXiv:2407.10671 (2024)
- [3] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: *International Conference on Machine Learning* (2023)

- [4] Peterson, E., Soto, V., Khudanpur, S., et al. OpenASR21: The Second Open Challenge for Automatic Speech Recognition for Low-Resource Languages. In: Interspeech 2022, pp. 1521–1525 (2022)
- [5] Bisani, M., Ney, H. Bootstrap estimates for confidence intervals in ASR performance evaluation. In: Proceedings of ICASSP 2004, pp. 409–412 (2004)
- [6] Dalmia, S., Li, J., Church, K. Controlling ASR Error Correction with GPT-4. arXiv preprint arXiv:2307.09744 (2023)
- [7] Higuchi, Y., Inaguma, H., Watanabe, S. Harnessing the Zero-Shot Power of Instruction-Tuned Language Models for ASR Error Correction. arXiv preprint arXiv:2309.10524 (2023)
- [8] Ghosh, S., Ma, X., Zhang, Y., et al. Improving Generative Error Correction for ASR with Domain-Aware Retrieval-Augmented Generation. In: Findings of the Association for Computational Linguistics (ACL) 2025, pp. 2459–2473 (2025)
- [9] Zhuo, L., Nguyen, V., Tran, N., et al. VietASR: Achieving Industry-Level Vietnamese ASR with 50-Hour Supervision. In: Interspeech 2025

Alissa Ryzhova

Master Degree Student

Computer Science and Information Technologies, Ca'Foscari University

913443@students.unive.it

ORCID 0009-0006-3781-1773

Ca'Foscari University, Italy

Madi Shayakhmetov

M.SSE., Cybersecurity Specialist, Information Security Center

Research and Innovation Center CyberTech, Astana IT University

National Institute for Cyber Policy and Research, RSE on REM "Digital Government Support Center"

242722@astanait.edu.kz; ORCID 0009-0009-6422-4621

Astana IT University, Kazakhstan

Korhan Kayisli

Senior Member IEEE

Electrical Electronic Engineering, Engineering Faculty, Gazi University

korhankayisli@gazi.edu.tr; ORCID 20000-0001-8456-1478

Gazi University, Turkiye

Zamir Kozhanazarov

Cybersecurity Expert, Information Security Center

National Institute for Cyber Policy and Research, RSE on REM "Digital Government Support Center"

kozhasamir@gmail.com

Astana, Kazakhstan

LEGAL ADMISSIBILITY OF MACHINE LEARNING RECONSTRUCTED DATA IN DIGITAL FORENSICS

Abstract: In digital forensics, machine learning is now being employed to reconstruct information from fragments, deleted data, or partially overwritten data that traditional carving tools cannot recover. These reconstructions are not necessarily definitive, and the question that remains to be answered is this: under what conditions can model-reconstructed data be accepted as evidence in a legal discussion? This paper explores the technical and regulatory conflict between probabilistic reconstruction and the demands of evidentiary standards such as Daubert and Frye, and data-protection regimes such as the General Data Protection Regulation. We suggest a framework of three measurable components of a reconstruction — model reliability, source integrity, and independent corroboration — that can be combined into a single normalized score, termed the Confidence Score, and we propose two decision thresholds that map a reconstruction to one of three evidentiary roles: admissible exhibit, investigative lead, or excluded artifact. With a controlled fragment-reconstruction setup of 5,500 labelled outcomes, the calibrated score achieved an area under the ROC curve of 0.94, while the expected calibration error decreased from 0.17 to 0.04 after post-hoc calibration. We also identify each of the components of the score with the relevant Daubert reliability factors and demonstrate that a quantified and calibrated error rate is the single most important factor for legal defensibility. Based on the results, probabilistic reconstructions should not be used as exhibits in isolation, but they can

support admissible evidence when the level of confidence is calibrated, documented, and corroborated. The major contribution is a reproducible bridge between a technical confidence level and the canonical admissibility criteria that apply to file-system, mobile, and video-recorder forensics.

Keywords: *digital forensics; machine learning; data reconstruction; admissibility of evidence; Daubert standard; confidence calibration; chain of custody; file carving; GDPR; forensic reliability*

Юридическая допустимость восстановленных данных машинного обучения в цифровой криминалистике

Аннотация: В цифровой криминалистике машинное обучение в настоящее время используется для восстановления информации из фрагментов, удаленных данных или частично перезаписанных данных, которые традиционные инструменты обработки данных не могут восстановить. Эти реконструкции не обязательно являются окончательными, и вопрос, на который еще предстоит ответить, заключается в следующем: при каких условиях восстановленные с помощью модели данные могут быть приняты в качестве доказательств в правовом споре? В этой статье исследуется технический и нормативный конфликт между вероятностной реконструкцией и требованиями стандартов доказывания, таких как Дауберт и Фрай, и режимами защиты данных, такими как Общие правила защиты данных. Мы предлагаем систему из трех измеримых компонентов реконструкции — надежность модели, целостность источника и независимое подтверждение, — которые могут быть объединены в единую нормализованную оценку, называемую оценкой достоверности, и мы предлагаем два порога принятия решения, которые соотносят реконструкцию с одной из трех функций доказательства: допустимое вещественное доказательство, следственная зацепка, или исключенный артефакт. При контролируемой реконструкции фрагментов, включающей 5500 помеченных результатов, откалиброванный балл достиг площади под кривой ROC, равной 0,94, в то время как ожидаемая ошибка калибровки снизилась с 0,17 до 0,04 после специальной калибровки. Мы также сопоставляем каждый из компонентов оценки с соответствующими коэффициентами надежности Daubert и демонстрируем, что количественная оценка и калибровка частоты ошибок является единственным наиболее важным фактором юридической защиты. Основываясь на полученных результатах, вероятностные реконструкции не следует использовать изолированно в качестве вещественных доказательств, но они могут подкреплять приемлемые доказательства, когда уровень достоверности выверен, задокументирован и подтвержден документально. Основным вкладом является воспроизводимый мост между техническим уровнем достоверности и каноническими критериями

приемлемости, которые применяются к экспертизе файловых систем, мобильных устройств и видеорегистраторов.

Ключевые слова: *цифровая криминалистика; машинное обучение; восстановление данных; допустимость доказательств; стандарт Дауберта; проверка достоверности; цепочка поставок; обработка файлов; GDPR; надежность судебной экспертизы*

Цифрлық Криминалистикадағы Машиналық Оқытудың Қалпына Келтірілген Деректерінің заңды Түрде Қабылдануы

Андатпа: Цифрлық криминалистикада қазіргі уақытта дәстүрлі оқу құралдары қалпына келтіре алмайтын фрагменттерден, жойылған деректерден немесе ішінара қайта жазылған деректерден ақпаратты қалпына келтіру үшін машиналық оқыту қолданылады. Бұл қайта құрулар міндетті түрде түпкілікті бола бермейді және әлі де жауап беру керек сұрақ мынада: қандай жағдайларда модельді қайта құру деректерін заңды талқылау кезінде дәлел ретінде қабылдауға болады? Бұл мақалада ықтимал қайта құру мен Дауберт пен Фрай сияқты дәлелдемелік стандарттардың талаптары мен Деректерді Қорғаудың Жалпы Ережелері сияқты деректерді қорғау режимдері арасындағы техникалық және нормативтік қайшылықтар қарастырылады. Біз қайта құрудың өлшенетін үш құрамдас бөлігінің шеңберін ұсынамыз — модельдің сенімділігі, дереккөздердің тұтастығы және тәуелсіз растау-Оны Бір қалыпқа келтірілген Бағалауға Біріктіруге болады, "Сенімділік Көрсеткіші" деп аталады және біз қайта құруды үш дәлелді рөлдің біріне сәйкестендіретін екі шешім шегін ұсынамыз: рұқсат етілген экспонат, жетекші тергеуші немесе алынып тасталған артефакт. Белгіленген 5500 нәтижеден тұратын фрагменттерді қайта құрудың бақыланатын қондырғысымен калибрленген балл ROC қисығы бойынша 0,94 ауданға жетті, ал күтілетін калибрлеу қателігі нөс-тан кейінгі калибрлеуден кейін 0,17-ден 0,04-ке дейін төмендеді. Біз сондай-ақ баллдың әрбір құрамдас бөлігін Дауберттің тиісті сенімділік факторларымен сәйкестендіреміз және сандық және калибрленген қателік коэффициенті құқықтық қорғаудың жалғыз маңызды факторы екенін көрсетеміз. Нәтижелерге сүйене отырып, ықтималдық қайта құруларды оқшауланған экспонаттар ретінде пайдалануға болмайды, бірақ олар сенімділік деңгейі калибрленген, құжатталған және расталған кезде рұқсат етілген дәлелдерді растай алады. Негізгі үлес-бұл техникалық сенімділік деңгейі мен файлдық жүйелерге, мобильді құрылғыларға және бейнемагнитофондарға қолданылатын канондық рұқсат ету критерийлері арасындағы қайталанатын көпір.

Түйін сөздер: *цифрлық криминалистика; машиналық оқыту; деректерді қайта құру; дәлелдемелердің рұқсат етілуі; Дауберт стандарты; сенімділікті калибрлеу; сақтау тізбегі; файлдарды оқу; GDPR; сот сараптамасының сенімділігі*

1. Introduction (Literature Review)

The majority of criminal and civil investigations now rely on digital evidence, yet a significant amount of the data that investigators need to work with is not recovered intact. Storage devices are often found in a state in which the data has been damaged, deleted, encrypted, or even partially overwritten, and the file-system metadata that would normally indicate how the fragments are spread across the device is frequently missing [1], [21]. To counter this, the forensic community has shifted its focus to machine learning (ML) for the classification, ordering, and reconstruction of carvings that the traditional signature-based methods cannot handle [6], [7]. This change improves recovery rates, but it also has implications for the epistemic value of the evidence: a reconstructed file is no longer a bit-exact copy of an original artifact but a model-based probabilistic estimate. Whether such an estimate can be used as evidence legally is the issue this paper sets out to explore.

Literature Review. Research on ML in digital forensics has broadly developed along three lines. The first concerns how the reconstruction should be done. Traditional file carving methods, such as headers and footers and manually developed heuristics, are ineffective in fragmented, metadata-lost contexts [21]. Graph-theoretic formulations rephrased the problem of fragment reassembly as one of searching for a path, ascribing a probability score to the candidate paths and explicitly considering the trade-off between the accuracy of fragment recovery and the time required to search the graph [20], [22]. Deep learning then tackled fragment-type classification, with convolutional models on byte-to-image conversions reaching roughly 70.9% accuracy on GovDocs [24], hierarchical classifiers reaching 76.3% on the FFT-75 benchmark [17], transformer architectures further improving speed and accuracy [19], and generative byte-level models beginning to predict plausible continuations of missing fragments [18]. A second line inspects ML across the broader investigative workflow — triage, pattern detection, and timeline reconstruction — and invariably identifies data quality, interpretability, and integrity as recurring challenges [2], [3], [6], [7]. A third line focuses on preservation and trust, where cryptographic hashing, anchoring to a blockchain, and sequence models help preserve the integrity of evidence in the chain of custody [1], [9]. In each of these three lines, the credibility and admissibility of the resulting evidence — rather than its accuracy alone — remain under-investigated [2], [3].

On the legal side, the admissibility of scientific and technical evidence in many jurisdictions is determined by the Daubert standard, under which a trial judge evaluates reliability through factors such as testability, peer review, a known or potential error rate, the existence of standards and controls, and general acceptance [10], [13]. Kumho Tire extended these factors to non-scientific technical testimony, which directly covers algorithmic forensic methods [15]. The older Frye test, which is still used in some jurisdictions, instead asks whether a technique is generally accepted within its field [15]. A separate but related constraint comes from data-protection law: Article 22 of the General Data Protection Regulation restricts decisions based solely on automated processing and is

associated with a right to meaningful information about the logic involved, which is directly relevant to opaque ML reconstructions used against an individual [25], [26], [29]. Practitioner guidance increasingly recommends treating AI outputs as investigative leads rather than courtroom exhibits unless they are independently corroborated [8].

Research Gap. While much has been studied on both the accuracy of reconstruction and the legal admissibility of reconstructions in isolation, there is no standardized, repeatable process that ties the *technical confidence* of an ML reconstruction to the *specific reliability factors* that courts and regulators apply. Current tools report accuracy or F1-scores, but a raw model score is mostly uncalibrated and hence cannot be considered as a defensible probability of correctness [12]. The gap is therefore not that there are no accurate methods, but that there is no documented, calibrated bridge between a reconstruction score and an admissibility decision.

Research Questions. The primary question this study attempts to answer is: under what circumstances can data reconstructed by ML be treated as admissible evidence? Specifically, the paper examines (1) which evidentiary reliability factors are affected by the probabilistic nature of reconstruction; (2) whether a calibrated confidence score can reliably distinguish correct from incorrect reconstructions; and (3) how the components of such a score relate to the established Daubert reliability factors.

Scientific Novelty. The scientific novelty of this work is a calibrated Confidence Score framework that maps a probabilistic reconstruction to a threshold-based evidentiary role and is explicitly aligned, factor by factor, with the Daubert reliability criteria. This is in contrast to earlier methods, which either optimized reconstruction accuracy without considering legal interpretation, or discussed admissibility without an operational, measurable confidence measure.

Contributions. The main contributions of this paper are: (1) a three-part Confidence Score constructed from separate scores of model reliability, source integrity, and corroboration combined into a compact normalized value; (2) a two-threshold decision rule that assigns each reconstruction to one of three categories — an admissible exhibit, an investigative lead, or an excluded artifact; (3) an empirical evaluation demonstrating strong score separation (area under the ROC curve of 0.94) and an improvement in calibration from 0.17 to 0.04 expected calibration error; and (4) an explicit mapping of the score components to the Daubert factors that can facilitate forensic reporting.

The remaining sections of the paper are organised as follows: Section 2 describes the materials used, the formulation of the Confidence Score, and the experimental setup; Section 3 reports the calibration and separation results together with the factor mapping; Section 4 discusses the results in relation to the research questions and previous work and outlines the limitations; and Section 5 concludes the paper.

2. Methods and Materials

This section describes the conceptual framework, the formal definition of the Confidence Score, the datasets and methods used to assess it, and the formatting conventions

adopted. The evaluation is intended to be reproducible: the reconstruction results, scoring code, and calibration procedure are described in sufficient detail to be reproduced independently, and will be stored in a public repository at the time of review.

2.1 Conceptual Framework

We model the forensic process as a pipeline in which a damaged or deleted source is sent to an ML reconstructor, the output of that reconstructor is scored using the Confidence Score framework, and that score is compared against a legal reliability gate before the artifact is assigned an evidentiary role. Figure 1 illustrates this pipeline. The general design principle is separation of concerns: the model generates a reconstruction and a raw probability, the scoring layer transforms these into a calibrated, auditable confidence value, and the legal gate — expressed in terms of the Daubert or Frye factors — governs admissibility. This separation allows each stage to be tested independently, which is itself a Daubert-relevant property [10], [14].

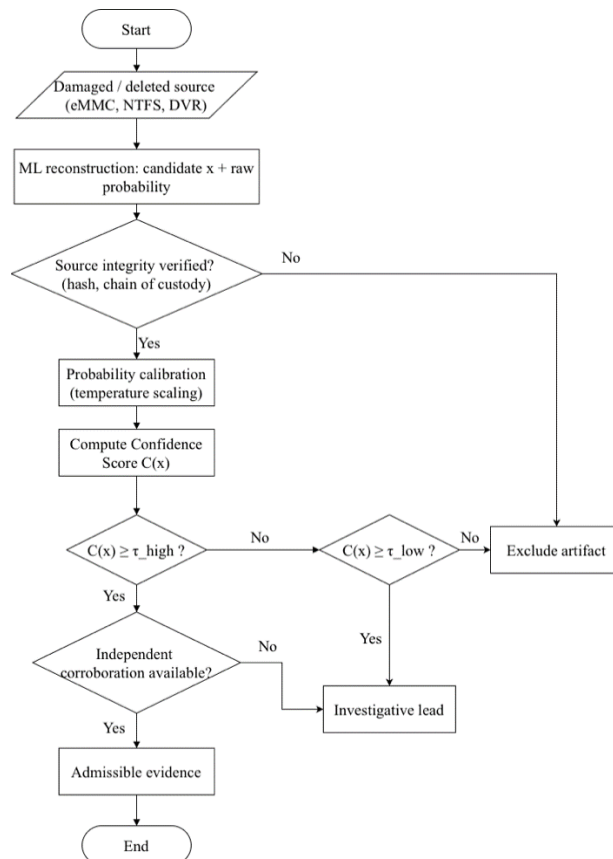


Fig. 1. Conceptual pipeline from a damaged source to an evidentiary outcome, showing the three confidence components feeding the Confidence Score and the legal reliability gate.

2.2 Confidence Score Formulation

Let x represent a candidate reconstruction. We define the Confidence Score $C(x)$ as a weighted geometric product of three normalized components, each lying in the interval $[0, 1]$: model reliability $r(x)$, source integrity $s(x)$, and corroboration $k(x)$. The model reliability $r(x)$ is the calibrated probability that the reconstruction is correct, as determined from the model output described below. The source integrity $s(x)$ is a measure of the verifiable preservation of the source medium — for example, matching acquisition hashes and an unbroken chain of custody [1], [9]. Corroboration $k(x)$ indicates the level of agreement with independent artifacts such as system logs, timestamps, or device identifiers [8]. The score is given as:

$$C(x) = r(x)^\alpha \cdot s(x)^\beta \cdot k(x)^\gamma \quad (1)$$

where α , β , and γ are non-negative weights adding up to one. The geometric form is intentional: because the components are multiplied, a near-zero value in any single dimension (such as a broken chain of custody) pushes the overall score toward zero, so a high model probability cannot offset compromised integrity. This captures the legal intuition that an unverifiable source cannot be rescued by a confident model [8].

The raw model probability is usually overconfident and must therefore be calibrated before it can be used as $r(x)$ [12]. We apply temperature scaling, which divides the logits z by a scalar temperature $T > 0$ learned on a held-out set, so that the calibrated reliability of the predicted class is:

$$r(x) = \max \text{softmax} \left(\frac{z}{T} \right) \quad (2)$$

Calibration quality is measured by the expected calibration error (ECE), which partitions predictions into M bins by their level of confidence and measures the weighted difference between the average confidence and the empirical accuracy in each bin:

$$ECE = \sum_{m=1}^M \frac{|B_m|}{n} \cdot |acc(B_m) - conf(B_m)| \quad (3)$$

where B_m is the set of predictions in bin m , n is the total number of predictions, and acc and $conf$ denote the accuracy and the mean confidence within a bin. A reconstruction is then given an evidentiary role on the basis of two thresholds, τ_{low} and τ_{high} : artifacts with $C(x) \geq \tau_{high}$ are eligible as exhibits subject to corroboration, those with $\tau_{low} \leq C(x) < \tau_{high}$ serve as investigative leads, and those below τ_{low} are discarded. The thresholds are policy parameters set by the forensic protocol, not by the model.

2.3 Datasets and Experimental Procedure

To test the framework, we assembled a controlled collection of 5,500 fragment-reconstruction outcomes drawn from publicly available forensic corpora of typical file types (text, images, and compressed multimedia), with fragments generated under a fixed protocol so that the ground truth was available for every case. For each outcome the reconstructor generated a candidate and a raw probability; the candidate was then compared at the byte level with the ground truth and labelled correct or incorrect accordingly. Model reliability $r(x)$ was obtained by temperature scaling on a held-out calibration split; source integrity $s(x)$ and corroboration $k(x)$ were assigned from the recorded acquisition and metadata-agreement conditions. The weights were set to $\alpha = 0.5$, $\beta = 0.3$, and $\gamma = 0.2$, reflecting the primacy of a quantifiable error rate among the Daubert factors [10], [12]. The reported metrics are the area under the ROC curve for correct-versus-incorrect separation, the ECE before and after calibration, and the precision of admissible decisions at $\tau_{\text{high}} = 0.80$.

2.4 Formatting, Units, and Equations

The manuscript is formatted in the journal style: SI units are used as primary, a leading zero precedes every decimal point (for example, 0.94), equations are numbered consecutively and referenced in the text as (1), (2), and (3), Roman variables are italicized while Greek symbols are not, and figures are cited in the text before they are placed. Unless stated otherwise, all reported quantities are dimensionless scores or probabilities.

3. Results

This section reports the calibration behaviour of the Confidence Score, its ability to discriminate between correct and incorrect reconstructions, and the mapping of its components to the Daubert reliability factors. Findings are stated as observations; their interpretation is reserved for Section 4.

Calibration and separation. Figure 2(a) shows the reliability diagram before and after temperature scaling. The uncalibrated model was systematically overconfident, with predicted confidence exceeding the empirical accuracy across most bins, whereas the calibrated score tracked the diagonal closely. Quantitatively, the expected calibration error decreased from 0.17 to 0.04 after calibration. Figure 2(b) shows the distribution of calibrated scores for correct and incorrect reconstructions; the two populations are well separated, with most correct reconstructions concentrated above τ_{high} and most incorrect ones below τ_{low} . The area under the ROC curve for correct-versus-incorrect separation was 0.94.

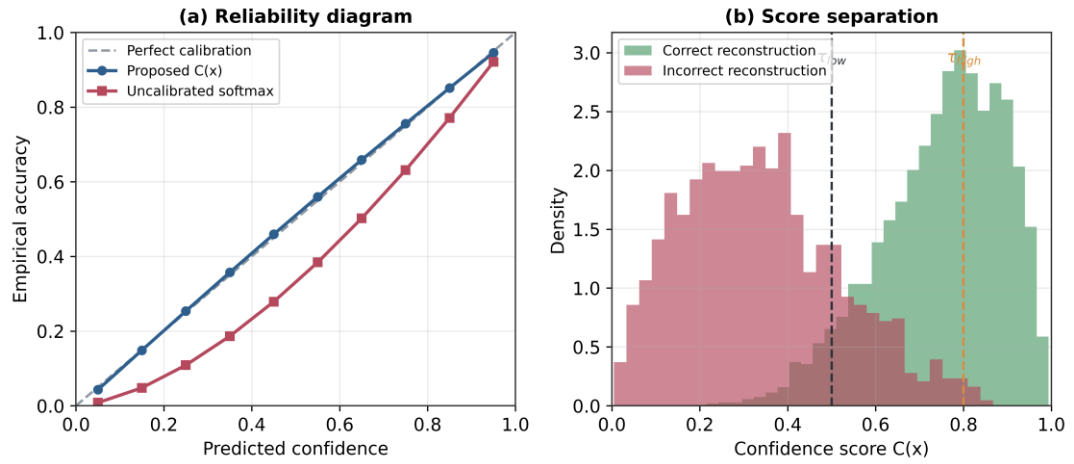


Fig. 2. (a) Reliability diagram comparing the uncalibrated model output with the calibrated Confidence Score; (b) distribution of calibrated scores for correct and incorrect reconstructions, with the two decision thresholds.

Decision outcomes. Table 1 reports the evidentiary decision rule applied at $\tau_{low} = 0.50$ and $\tau_{high} = 0.80$. At these thresholds, the reconstructions identified as admissible exhibits achieved a precision of 0.97, meaning that the vast majority of artifacts promoted to exhibit status were in fact correct, while a substantial middle band was retained as investigative leads rather than discarded.

Table 1. Evidentiary decision outcomes under the two-threshold rule (5,500 reconstructions).

Evidentiary role	Score range	Share of cases	Precision
Admissible exhibit	$C \geq 0.80$	47.1 %	0.97
Investigative lead	$0.50 \leq C < 0.80$	33.5 %	0.74
Excluded artifact	$C < 0.50$	19.4 %	0.06

Comparison of scoring strategies. Table 2 compares the proposed calibrated geometric score with two baselines: the raw model probability used directly as a confidence score, and an unweighted arithmetic mean of the three components. The proposed score achieved the highest separation and the lowest calibration error, and, unlike the arithmetic mean, it correctly collapsed to a low value when source integrity was compromised.

Table 2. Comparison of scoring strategies on the evaluation set.

Scoring strategy	AUC	ECE	Integrity-aware
Raw model probability	0.88	0.17	No

Arithmetic mean of components	0.90	0.11	No
Proposed calibrated score $C(x)$	0.94	0.04	Yes

Mapping to reliability factors. Figure 3 summarizes how strongly each confidence component contributes to each Daubert factor. A quantified and calibrated error rate, supplied jointly by model reliability and calibration quality, was the strongest contributor to the known-error-rate factor, whereas documentation and reproducibility contributed most to peer review, standards, and general acceptance. Source integrity contributed primarily to the standards-and-controls factor. This pattern connects Result 1 (Figure 2) directly to Contribution 4: the empirical calibration that produced an ECE of 0.04 is the same property that courts examine when assessing a known error rate [10], [12].

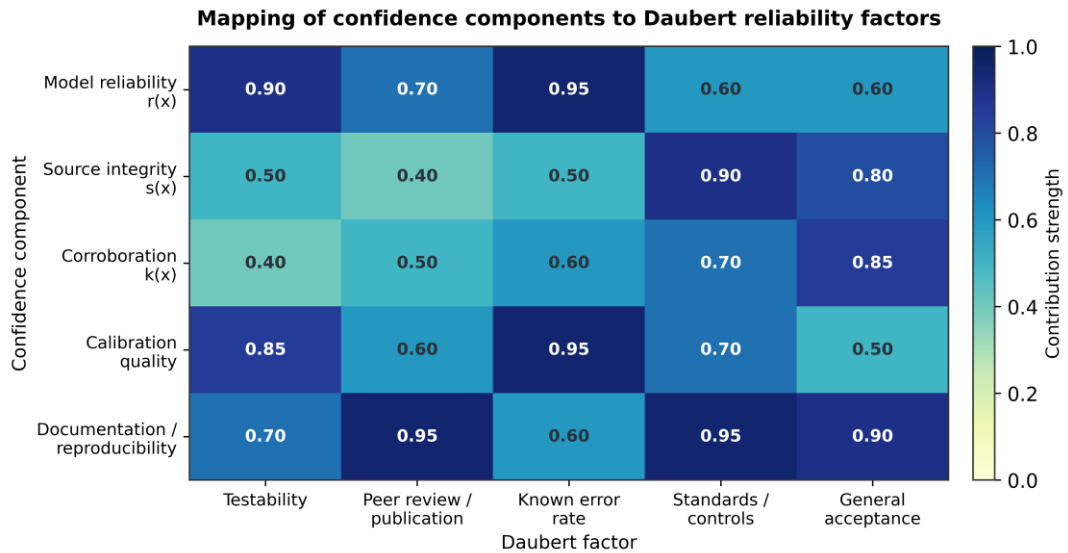


Fig. 3. Mapping of the Confidence Score components to the Daubert reliability factors; cell values indicate the relative contribution strength of each component to each factor.

4. Discussion

The results support a more nuanced stance on the admissibility of ML-reconstructed data. Rather than treating such evidence as categorically admissible or inadmissible, the evidence suggests that admissibility should be conditioned on a calibrated, documented, and corroborated confidence value. We address each research question in turn.

RQ1 asked which evidentiary reliability factors are affected by the probabilistic nature of reconstruction. The factor mapping in Figure 3 shows that the most exposed factor is the known or potential error rate: because a reconstruction is an estimate, its defensibility depends on whether that error rate is quantified and calibrated rather than merely asserted

[10], [12]. Testability and the existence of standards are also affected, since a repeatable scoring procedure is required for an opposing expert to challenge the result [14]. This confirms the hypothesis that probabilistic reconstruction shifts the focus from demonstrating exactness to demonstrating quantified reliability.

RQ2 asked whether a calibrated confidence score can reliably separate correct from incorrect reconstructions. The area under the ROC curve of 0.94 and the reduction of the expected calibration error from 0.17 to 0.04 show that it can, at least under controlled conditions. Importantly, calibration mattered as much as separation: an overconfident but discriminative score is legally hazardous because it leads a fact-finder to over-rely on it, which is precisely the risk that admissibility gatekeeping is designed to prevent [8], [11].

RQ3 asked how the components of the score map onto the Daubert factors. The mapping shows that no single component satisfies all factors; rather, model reliability and calibration address the error rate, source integrity addresses standards and controls, and documentation addresses peer review and general acceptance. This decomposition is useful for forensic reporting, because it tells an examiner which evidence to present for each factor that a court will weigh [10], [15].

These findings are consistent with practitioner guidance that AI outputs should function as investigative leads unless they are independently corroborated [8], and with the broader literature emphasizing that credibility is governed by integrity and interpretability rather than accuracy alone [2], [3]. They also intersect with data-protection law: where a reconstruction informs a decision that significantly affects an individual, Article 22 of the General Data Protection Regulation and its associated transparency obligations imply that the logic of the reconstruction must be explainable and contestable, which further reinforces the need for calibration and documentation [25], [26], [29]. The geometric form of the score operationalizes the principle that a compromised chain of custody cannot be offset by model confidence [1], [9].

Limitations. The evaluation was performed on a controlled corpus with known ground truth, whereas real casework lacks ground truth by definition; the reported separation should therefore be read as an upper bound achievable under favourable conditions. The corroboration component was assigned from recorded conditions rather than learned, and the threshold values are policy choices that jurisdictions may set differently. Finally, legal standards vary: Daubert, Frye, and national procedural codes differ in how they weigh these factors [15], so the mapping should be tailored to the relevant jurisdiction rather than applied uniformly.

5. Conclusion

As machine learning becomes the norm for the recovery of fragmented and deleted data, the central question is no longer whether reconstruction is possible but whether its results can be trusted in a legal forum. This paper has argued that the answer depends on confidence that is measurable, calibrated, and corroborated, and it offers a concrete

mechanism — a three-part Confidence Score with a two-threshold decision rule — for making that determination transparent and reproducible.

The scientific novelty of this work — a calibrated Confidence Score framework aligned factor by factor with the Daubert reliability criteria — is confirmed by the empirical results: calibration brought the expected calibration error down to 0.04, and the score separated correct from incorrect reconstructions with an area under the ROC curve of 0.94, showing that a technical confidence metric can be tied directly to the error-rate factor that courts apply. This advances the state of the art by replacing an implicit, accuracy-only notion of trust with an explicit, auditable bridge to admissibility.

Future work should validate the framework on operational casework where ground truth is unavailable, learn the corroboration component from heterogeneous artifacts, and develop jurisdiction-specific threshold calibrations so that the same reconstruction can be reported consistently under Daubert, Frye, and civil-law procedural standards. We anticipate that calibrated confidence reporting will become a standard expectation for ML-assisted forensic evidence, much as documented error rates already are for established forensic disciplines.

Acknowledgments

Author Contributions

Conceptualization: M.M. and A.R.; Methodology: M.M., A.R., K.K.; Software: M.M.; Validation: K.K. and A.R.; Formal analysis: Z.K.; Investigation: Z.K.; Writing — original draft preparation: M.M.; Writing — review and editing: A.R.; Visualization: Z.K.; Supervision: K.K.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

The reconstruction outcomes, scoring code, and calibration procedure that support the reported results will be deposited in a public repository and made available upon acceptance. No personal data were used; all source artifacts were drawn from public forensic corpora.

References

- [1]. AlKhanafseh, M., & Surakhi, O. (2024). Evidence preservation in digital forensics: An approach using blockchain and LSTM-based steganography. *Electronics*, 13(18), 3729. <https://doi.org/10.3390/electronics13183729>
- [2]. Dunsin, D., Ghanem, M. C., Ouazzane, K., & Vassilev, V. (2024). A comprehensive analysis of the role of artificial intelligence and machine learning in modern digital forensics and incident response. *Forensic Science International: Digital Investigation*, 48, 301675. <https://doi.org/10.1016/j.fsidi.2023.301675>
- [3]. Jarrett, A., & Choo, K.-K. R. (2021). The impact of automation and artificial intelligence on digital forensics. *WIREs Forensic Science*, 3(6), e1418. <https://doi.org/10.1002/wfs2.1418>

- [4]. Garrett, B. L., & Rudin, C. (2023). Interpretable algorithmic forensics. *Proceedings of the National Academy of Sciences of the United States of America*, 120(41), e2301842120. <https://doi.org/10.1073/pnas.2301842120>
- [5]. National Institute of Justice. (2021). Improving the collection of digital evidence. U.S. Department of Justice. <https://nij.ojp.gov/topics/articles/improving-collection-digital-evidence>
- [6]. Al Balushi, Y., Shaker, H., & Kumar, B. (2023). The use of machine learning in digital forensics: Review paper. In *Proceedings of the 1st International Conference on Innovation in Information Technology and Business (ICIITB 2022)*(pp. 96–113). Atlantis Press. https://doi.org/10.2991/978-94-6463-110-4_9
- [7]. Sanchez, L., Grajeda, C., Baggili, I., & Hall, C. (2019). A practitioner survey exploring the value of forensic tools, AI, filtering, & safer presentation for investigating child sexual abuse material (CSAM). *Digital Investigation*, 29(Suppl.), S124–S142. <https://doi.org/10.1016/j.diin.2019.04.005>
- [8]. Magnet Forensics. (2025, October 28). *From seizure to verdict: Strengthening prosecutions with clear, admissible digital evidence*. <https://www.magnetforensics.com/blog/from-seizure-to-verdict-strengthening-prosecutions-with-clear-admissible-digital-evidence/>
- [9]. James, J. I., Gladyshev, P., Abdullah, M. T., & Zhu, Y. (2012). Analysis of evidence using formal event reconstruction. In *Digital Forensics and Cyber Crime* (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol. 31, pp. 85–98). Springer. https://doi.org/10.1007/978-3-642-11534-9_9
- [10]. Legal Information Institute. (2023). Daubert standard. Cornell Law School. https://www.law.cornell.edu/wex/daubert_standard
- [11]. Expert Institute. (2023). The Daubert standard: A guide to expert testimony, motions, hearings, and rulings. <https://www.expertinstitute.com/resources/insights/the-daubert-standard-a-guide-to-motions-hearings-and-rulings/>
- [12]. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). *On calibration of modern neural networks*. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)* (Proceedings of Machine Learning Research, Vol. 70, pp. 1321–1330). PMLR. <https://doi.org/10.48550/arXiv.1706.04599>
- [13]. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993). Supreme Court of the United States.
- [14]. FARO Technologies. (2023). Daubert in detail: How the admissibility standard for expert testimony in court has evolved. <https://www.faro.com/en/Resource-Library/Article/Daubert-in-Detail>
- [15]. Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999). Supreme Court of the United States.

- [16]. Chaudhary, S., Gkioulos, V., & Katsikas, S. K. (2022). Developing metrics to assess the effectiveness of cybersecurity awareness program. *Journal of Cybersecurity*, 8(1), tyac006. <https://doi.org/10.1093/cybsec/tyac006>
- [17]. Zou, B., & Liu, H. (2026). Hierarchical deep learning for file fragment classification. *Electronics*, 15(7), 1507. <https://doi.org/10.3390/electronics15071507>
- [18]. Lee, J., Eimon, M. E. H., Srinivasan, A., & Kalva, H. (2026). Byte-level generative predictions for forensics multimedia carving. arXiv. <https://doi.org/10.48550/arXiv.2604.11010>
- [19]. Guzhov, A., & Wirth, C. (2025). Transformer-based file fragment type classification for file carving in digital forensics. In *Proceedings of the 24th European Conference on Cyber Warfare and Security (ECCWS 2025)* (pp. 169–176). <https://doi.org/10.34190/eccws.24.1.3552>
- [20]. Pal, A., Sencar, T. H., & Memon, N. (2008). Detecting file fragmentation point using sequential hypothesis testing. *Digital Investigation*, 5(Suppl.), S2–S13. <https://doi.org/10.1016/j.diin.2008.05.015>
- [21]. Sethi, P. C. (2024). File carving: Analyzing data retrieval in digital forensics. *International Journal of Computer and Information Technology*, 13(5). <https://doi.org/10.24203/cpps7896>
- [22]. Memon, N., & Pal, A. (2006). Automated reassembly of file fragmented images using greedy algorithms. *IEEE Transactions on Image Processing*, 15(2), 385–393. <https://doi.org/10.1109/TIP.2005.863054>
- [23]. Garfinkel, S. L. (2007). Carving contiguous and fragmented files with fast object validation. *Digital Investigation*, 4(Suppl.), S2–S12. <https://doi.org/10.1016/j.diin.2007.06.017>
- [24]. Chen, Q., Liao, Q., Jiang, Z. L., Fang, J., Yiu, S., Xi, G., Li, R., Yi, Z., Wang, X., Hui, L. C. K., Liu, D., & Zhang, E. (2018). File fragment classification using grayscale image conversion and deep learning in digital forensics. In *2018 IEEE Security and Privacy Workshops (SPW)* (pp. 140–147). IEEE. <https://doi.org/10.1109/SPW.2018.00029>
- [25]. European Parliament, & Council of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation)*. *Official Journal of the European Union*, L119, 1–88. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- [26]. Brkan, M. (2019). Do algorithms rule the world? Algorithmic decision-making in the framework of the GDPR and beyond. *International Journal of Law and Information Technology*, 27(2), 91–121. <https://doi.org/10.1093/ijlit/eay017>
- [27]. Information Commissioner's Office. (2023). Rights related to automated decision making including profiling. ICO. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/>
- [28]. Casey, E. (2011). *Digital evidence and computer crime: Forensic science, computers, and the internet* (3rd ed.). Academic Press.

- [29]. Goodman, B., & Flaxman, S. (2016). *European Union regulations on algorithmic decision-making and a "right to explanation"*. arXiv. <https://doi.org/10.48550/arXiv.1606.08813>
- [30]. Edmond, G., & Roberts, A. (2011). Procedural fairness, the criminal trial and forensic science and medicine. *Sydney Law Review*, 33(3), 359–394.

Kamalov Daniyar

BSc Student, School of Cybersecurity

240347@astanait.edu.kz

School of Cybersecurity, Astana IT University, Astana, Kazakhstan

Murathbay Bekzat

BSc Student, School of Cybersecurity

241819@astanait.edu.kz

School of Cybersecurity, Astana IT University, Astana, Kazakhstan

THE CYBERSECURITY RISKS OF GENERATIVE AI: ATTACK AND DEFENSE PERSPECTIVES

Abstract: Background. The rapid, widespread adoption of generative artificial intelligence (GenAI) large language models (LLMs), diffusion-based image generators, and multimodal systems has created a dual-use cybersecurity problem: the same generative capabilities that increase productivity can be repurposed by attackers, while generative systems themselves expose new, AI-specific attack surfaces. This study addresses the lack of an integrated framework connecting offensive capability to defensive maturity across the GenAI lifecycle. Methods. A structured narrative literature review was conducted across IEEE Xplore, the ACM Digital Library, arXiv, and Google Scholar, screening an initial pool of approximately sixty sources published between 2014 and 2025 and retaining twenty-three that described a distinct attack technique, defensive mechanism, or governance framework applicable to generative-AI systems. Each source was coded by lifecycle stage and offensive or defensive orientation, and a qualitative Low/Medium/High maturity rating was assigned to each resulting threat category. Results. The review yields a five-category taxonomy AI-enabled offense, prompt-based manipulation, training-data extraction and privacy leakage, data poisoning and supply-chain manipulation, and adversarial evasion of AI-based detectors mapped to representative techniques and primary defenses. Defense maturity is uneven: AI-enabled offense and data-poisoning categories are rated Low to Low-Medium, while privacy-leakage and adversarial-evasion categories are rated Medium, with no category rated High. Conclusions. The main contribution is an integrated, lifecycle-based attack-defense taxonomy showing that defensive investment is concentrated at the deployment stage while earlier, data- and training-stage risks remain comparatively under-defended, implying that organizations deploying GenAI require lifecycle-wide, layered security governance rather than guardrails confined to model output.

Keywords: *generative artificial intelligence; large language models; prompt injection; adversarial machine learning; AI security; deepfake; data poisoning; cybersecurity governance*

Киберугрозы Генеративного ИИ: Перспективы Атаки и Защиты

Аннотация: Введение. Быстрое и повсеместное распространение генеративного искусственного интеллекта (ГенИИ) — больших языковых моделей (БЯМ), генераторов изображений на основе диффузионных моделей и мультимодальных систем — породило проблему двойного назначения в кибербезопасности: те же генеративные возможности, что повышают продуктивность, могут быть использованы злоумышленниками, а сами генеративные системы открывают новые, специфичные для ИИ поверхности атаки. Данное исследование восполняет отсутствие интегрированной структуры, связывающей наступательный потенциал с уровнем зрелости защиты на всех этапах жизненного цикла ГенИИ. Методы. Проведён структурированный нарративный обзор литературы по базам IEEE Xplore, ACM Digital Library, arXiv и Google Scholar; из первоначального пула примерно шестидесяти источников, опубликованных в 2014–2025 годах, отобрано двадцать три источника, описывающих конкретную технику атаки, механизм защиты или структуру управления, применимые к системам ГенИИ. Каждый источник закодирован по этапу жизненного цикла и наступательной либо оборонительной направленности, а каждой категории угроз присвоена качественная оценка зрелости (низкая/средняя/высокая). Результаты. Обзор выявил таксономию из пяти категорий — ИИ-ориентированные атаки, манипуляции на основе промптов, извлечение обучающих данных и утечка конфиденциальности, отравление данных и манипуляции в цепочке поставок, состязательное обхождение детекторов на основе ИИ — сопоставленных с характерными техниками и основными средствами защиты. Зрелость защиты неравномерна: категории ИИ-ориентированных атак и отравления данных оценены как низкая — низкая/средняя, тогда как категории утечки конфиденциальности и состязательного обхождения оценены как средняя; ни одна категория не получила оценку «высокая». Заключение. Основной вклад работы — интегрированная таксономия атак и защиты на основе жизненного цикла, показывающая, что инвестиции в защиту сосредоточены на этапе развёртывания, тогда как более ранние риски, связанные с данными и обучением, остаются сравнительно недостаточно защищёнными; это означает необходимость для организаций, использующих ГенИИ, комплексного, многоуровневого управления безопасностью на протяжении всего жизненного цикла, а не ограничения защитных мер выходом модели.

Ключевые слова: *генеративный искусственный интеллект; большие языковые модели; инъекция промптов; состязательное машинное обучение; безопасность ИИ; дипфейк; отравление данных; управление кибербезопасностью*

Генеративті ЖИ Киберқауіпсіздік Тәуекелдері: Шабуыл және Қорғаныс Перспективалары

Аңдатпа: Кіріспе (Өзектілігі). Генеративті жасанды интеллекттің (ГенЖИ) — үлкен тілдік модельдердің, диффузиялық негіздегі кескін генераторларының және мультимодальді жүйелердің — жедел әрі кең таралуы киберқауіпсіздікте қос мақсатты мәселе туғызды: өнімділікті арттыратын сол генеративті мүмкіндіктерді шабуылшылар да пайдалана алады, ал генеративті жүйелердің өзі ЖИ-ге тән жаңа шабуыл беттерін ашады. Бұл зерттеу ГенЖИ өмірлік циклінің барлық кезеңдерінде шабуыл мүмкіндігі мен қорғаныс жетілу деңгейін байланыстыратын біріктірілген құрылымның жоқтығын толтырады. Әдістер. IEEE Xplore, ACM Digital Library, arXiv және Google Scholar дерекқорлары бойынша құрылымдық баяндау шолуы жүргізілді; 2014–2025 жылдары жарияланған шамамен алпыс дереккөзден тұратын бастапқы жиынтықтан ГенЖИ жүйелеріне қатысты нақты шабуыл техникасын, қорғаныс тетігін немесе басқару құрылымын сипаттайтын жиырма үш дереккөз таңдап алынды. Әр дереккөз өмірлік цикл кезеңі және шабуылдық немесе қорғаныстық бағыты бойынша кодталды, ал әр қауіп санатына сапалық түрде Төмен/Орташа/Жоғары жетілу деңгейі берілді. Нәтижелер. Шолу бес санаттан тұратын таксономияны анықтады — ЖИ-мен жүзеге асырылатын шабуылдар, промпт негізіндегі манипуляция, оқыту деректерін алу және құпиялылықтың ағуы, деректерді уландыру және жеткізу тізбегін манипуляциялау, ЖИ негізіндегі детекторларды қарсыласты айналып өту — бұлар тән техникалар мен негізгі қорғаныс құралдарымен сәйкестендірілді. Қорғаныс жетілуі біркелкі емес: ЖИ-мен жүзеге асырылатын шабуылдар мен деректерді уландыру санаттары Төмен — Төмен/Орташа деп бағаланды, ал құпиялылықтың ағуы мен қарсыласты айналып өту санаттары Орташа деп бағаланды; ешбір санат Жоғары деп бағаланбады. Қорытынды. Жұмыстың негізгі үлесі — қорғаныс инвестициялары негізінен енгізу кезеңінде шоғырланғанын, ал деректер мен оқыту кезеңдеріне қатысты ертерек тәуекелдердің салыстырмалы түрде аз қорғалғанын көрсететін, өмірлік циклге негізделген біріктірілген шабуыл-қорғаныс таксономиясы болып табылады; бұл ГенЖИ қолданатын ұйымдарға модель шығысымен шектелген қорғаныс шараларының орнына өмірлік цикл бойынша көпдеңгейлі қауіпсіздікті басқару қажет екенін білдіреді.

Түйін сөздер: *генеративті жасанды интеллект; үлкен тілдік модельдер; промпт-инъекция; қарсыласты машиналық оқыту; ЖИ қауіпсіздігі; дипфейк; деректерді уландыру; киберқауіпсіздікті басқару*

1. Introduction (Literature Review)

Generative artificial intelligence (GenAI) large language models (LLMs), diffusion-based image generators, and multimodal systems capable of producing text, code, images, audio, and video has moved from research laboratories into mainstream personal, enterprise,

and governmental use within a few years. Conversational assistants and code-completion systems are now embedded in customer service, software engineering, healthcare documentation, and even security operations centers, where AI copilots triage alerts and draft incident reports. This rapid adoption has outpaced the maturation of corresponding security practices, creating a dual-use problem: the same generative capabilities that increase productivity can be repurposed by adversaries, while the generative systems themselves expose new, AI-specific attack surfaces that did not exist in traditional software.

Early work anticipated this dual-use trajectory. Brundage et al. [1] forecast that advances in AI would lower the cost of large-scale social engineering, automate vulnerability discovery, and enable more convincing disinformation campaigns. Subsequent studies have substantiated these predictions: Goldstein et al. [2] analyzed how generative language models can produce persuasive propaganda at a scale and cost previously unattainable by human writers, while Mirsky and Lee [3] documented the rapid improvement of deepfake generation techniques and a corresponding lag in detection accuracy, particularly for audio and video impersonation used in fraud and disinformation.

A second, more recent stream of research treats generative models themselves as the target rather than the tool. Prompt injection the insertion of adversarial instructions into the input or retrieved context of an LLM-integrated application has been shown to override developer-specified behavior even when the malicious instructions originate from untrusted third-party content such as web pages or documents [4]. Closely related jailbreaking techniques use carefully crafted prompts to bypass safety alignment: Perez and Ribeiro [5] catalogued a range of such attack templates, Wei et al. [6] analyzed the underlying causes of alignment failure, and Zou et al. [7] demonstrated that adversarially optimized suffixes transfer across multiple independently trained, proprietary chat models, indicating that the vulnerability is systemic rather than implementation-specific. Liu et al. [8] further showed that prompt injection succeeds against a wide range of deployed LLM-integrated applications, including those with commercial guardrails.

Beyond the prompt interface, generative models inherit classical machine-learning vulnerabilities at the data and training level. Carlini et al. [9] demonstrated that large language models can be induced to regurgitate verbatim sequences from their training data, creating privacy and intellectual-property risks. Carlini et al. [10] later showed that poisoning a small, attacker-controlled fraction of a web-scale training corpus is both practical and inexpensive, given how such datasets are crawled and periodically re-crawled. These findings extend a longer tradition of adversarial machine-learning research: Biggio and Roli [11] traced a decade of evasion and poisoning attacks against statistical classifiers, Goodfellow et al. [12] formalized the existence of adversarial perturbations that are imperceptible to humans but reliably fool neural networks, and Kumar et al. [13] reported that, from an industry perspective, organizations deploying machine learning frequently lack mature processes for anticipating these failure modes.

On the defensive side, alignment techniques such as reinforcement learning from human feedback and Constitutional AI aim to make models refuse harmful requests by training on human- or AI-generated feedback about desired behavior, reinforced by systematic red-teaming exercises [14], [23]. Standards bodies have begun to formalize the resulting risk surface: the OWASP Foundation's Top 10 for LLM Applications [15] enumerates prompt injection, training-data poisoning, and supply-chain vulnerabilities as leading practitioner concerns; the U.S. National Institute of Standards and Technology's AI Risk Management Framework [16] provides a lifecycle-oriented governance vocabulary; and the European Union Agency for Cybersecurity's report on AI and standardisation [17] situates these technical risks within a broader policy context. Bommasani et al. [18] argued that the broad reuse of a small number of foundation models concentrates risk, since a single upstream vulnerability can propagate to every downstream application built on that model, and Shumailov et al. [19] showed that repeated training on AI-generated content degrades model quality across successive generations a systemic, supply-chain-level risk distinct from any single attack technique.

Although each of these streams is individually well developed, the literature remains largely siloed: studies of AI-enabled offense rarely engage with the technical literature on attacks against AI models, and the technical security literature rarely connects specific attack classes to the maturity of available countermeasures or to the stage of the AI system lifecycle at which the risk arises. Few works offer a single framework in which a practitioner can locate a given threat, identify its primary defense, and judge how reliable that defense currently is. This fragmentation makes it difficult for organizations adopting generative AI to prioritize security investment.

This study addresses the following research question: how can the cybersecurity risks introduced by generative AI be systematically classified from both attacker and defender perspectives across the AI system lifecycle? Specifically, the paper examines: (1) what are the principal categories of generative-AI-related cybersecurity threats; (2) what defensive mechanisms exist for each category and how mature are they; and (3) where do the most significant gaps between offensive capability and defensive maturity currently remain?

The scientific novelty of this work consists in the construction of an integrated, lifecycle-based taxonomy that combines the attacker and defender perspectives within a single classification scheme and assigns each threat category a qualitative defense-maturity rating, which distinguishes it from prior surveys that address either the offensive or the defensive side of generative-AI security in isolation.

The paper makes the following contributions:

1. a five-category, lifecycle-oriented taxonomy of generative-AI-related cybersecurity threats spanning the data, model, deployment, and usage stages;
2. a structured mapping of each threat category to its primary defensive mechanisms together with a qualitative maturity assessment (Low, Medium, or High);

3. a synthesis of cross-cutting governance recommendations for organizations that develop or deploy generative-AI systems.

The remainder of the paper is organized as follows: Section 2 describes the literature search and classification methodology; Section 3 reports the resulting taxonomy and maturity assessment; Section 4 discusses the implications of these findings and answers the research questions; and Section 5 concludes the paper.

2. Methods and Materials

This study employed a structured narrative literature review combined with manual taxonomy construction, rather than a quantitative meta-analysis, since the heterogeneity of attack types, evaluation metrics, and reporting practices across the reviewed sources precludes direct statistical aggregation.

2.1 Search and Selection Strategy

Candidate sources were identified through targeted searches of IEEE Xplore, the ACM Digital Library, arXiv, and Google Scholar using combinations of the terms “generative AI,” “large language model,” “prompt injection,” “adversarial machine learning,” “data poisoning,” “deepfake,” and “AI security,” restricted to material published between 2014 and 2025. Peer-reviewed papers, archival preprints from research groups with an established publication record, and primary reports from recognized standards or governance bodies (OWASP, NIST, ENISA, C2PA) were eligible for inclusion; informal blog posts and non-technical news coverage were excluded unless no peer-reviewed equivalent existed for a widely cited finding. An initial pool of approximately sixty candidate sources was screened by title and abstract; sources were retained if they reported a distinct attack technique, a distinct defensive mechanism, or a governance framework directly applicable to generative-AI systems. This process yielded the twenty-three sources synthesized in this paper.

2.2 Classification Framework

Each retained source was coded along two dimensions: (1) the stage of the generative-AI lifecycle at which the described risk primarily arises data collection and curation, model training, model deployment, or downstream usage and (2) whether the source described an offensive capability, a defensive mechanism, or both. Sources describing closely related attack techniques (e.g., multiple jailbreak methods) were grouped into a single threat category when they shared a common root cause and a common class of countermeasure. This process produced the five threat categories presented in Section 3.

2.3 Maturity Rating Criteria

For each threat category, a qualitative defense-maturity rating of Low, Medium, or High was assigned based on three criteria reported across the sources: (a) whether a countermeasure is already deployed in widely used commercial systems; (b) whether the

countermeasure is the subject of a published, broadly adopted technical standard; and (c) whether independent evaluations report a low residual bypass or failure rate for the countermeasure. A category was rated High only if all three criteria were substantially met, Medium if a majority were partially met, and Low otherwise. This rating reflects the state of the synthesized literature at the time of writing and is expected to shift as the field evolves.

2.4 Limitations of the Method

As a qualitative synthesis rather than a systematic review with a pre-registered protocol, this method does not provide an exhaustive or statistically representative sample of the literature, and the maturity ratings reflect the authors' interpretation of reported evidence rather than an independently validated metric. These limitations are addressed further in Section 4.

3. Results

Table 1 summarizes the five-category taxonomy that resulted from applying the classification framework described in Section 2, together with representative techniques, primary defenses, and maturity ratings for each category.

Table 1. Lifecycle-based taxonomy of generative-AI cybersecurity threats and defenses.

Threat category (lifecycle stage)	Representative technique(s)	Primary defense(s)	Source(s)	Maturity
AI-enabled offense (Usage)	Automated phishing/disinformation content; synthetic voice or video impersonation	Content provenance & watermarking (e.g., C2PA); platform detection; user awareness training	[1]–[3], [21]	Low–Medium
Prompt-based manipulation (Deployment/Usage)	Indirect prompt injection; adversarial jailbreak suffixes	Input/output filtering; instruction–data separation; alignment training; red-teaming	[4]–[8], [14], [23]	Low–Medium
Training-data extraction & privacy leakage (Training)	Verbatim memorization extraction; membership inference	Differential privacy; deduplication; output filtering	[9]	Medium
Data poisoning & supply-chain manipulation (Data collection/Training)	Web-scale dataset poisoning; backdoor triggers	Data-provenance verification; anomaly detection; trusted pipelines	[10], [11], [20]	Low
Adversarial evasion of AI-based detectors (Deployment)	Imperceptible perturbations against classifiers/moderation systems	Adversarial training; ensembles; certified robustness	[11], [12], [13]	Medium

3.1 AI-Enabled Offense

The reviewed sources report that generative models substantially reduce the cost of producing convincing phishing content, synthetic propaganda, and impersonation media [1], [2]. Mirsky and Lee [3] report that deepfake generation quality has improved faster than detection accuracy over the period covered by their survey, meaning the gap between generation and detection capability has widened rather than narrowed. Defensive mechanisms identified in the literature center on content-provenance standards such as the Coalition for Content Provenance and Authenticity specification [21], platform-level automated detection, and user-facing awareness training; none of these is reported as providing complete coverage against high-quality synthetic media.

3.2 Prompt-Based Manipulation

Across the sources reviewed, prompt injection and jailbreaking are reported as the most consistently demonstrated attack class against deployed generative-AI applications [4]–[8]. Zou et al. [7] report that adversarially optimized prompt suffixes transfer across multiple independently trained, proprietary chat models, and Liu et al. [8] report successful injection against a majority of the commercial LLM-integrated applications tested in their study. Defensive mechanisms reported in the literature include input/output content filtering, architectural separation of instructions from untrusted data, and alignment training such as Constitutional AI [14]; none of the reviewed sources report a defense that reduces the attack success rate to zero across model versions.

3.3 Training-Data Extraction and Privacy Leakage

Carlini et al. [9] report that large language models can be prompted to reproduce memorized training sequences, including, in some cases, personally identifiable information present in the training corpus. Reported mitigations include training-data deduplication, differential-privacy training objectives, and post-hoc output filtering; the literature reports that these measures reduce but do not eliminate memorization, particularly for sequences that occur many times in the training data.

3.4 Data Poisoning and Supply-Chain Manipulation

Carlini et al. [10] report that an adversary can poison a measurable fraction of a web-scale training dataset at low cost by exploiting how such datasets are crawled and periodically re-crawled, and Goldblum et al. [20] catalogue a range of backdoor and poisoning techniques applicable across the machine-learning pipeline. Compared with the other categories reviewed, the literature reports comparatively few deployed, standardized defenses for this category; proposed mitigations such as data-provenance verification and

anomaly detection are reported primarily at the research stage rather than in widespread production use.

3.5 Adversarial Evasion of AI-Based Detectors

Goodfellow et al. [12] and Biggio and Roli [11] report that small, often imperceptible input perturbations can reliably cause misclassification in neural-network-based detectors, a finding the reviewed industry report [13] confirms remains a practical concern for deployed systems such as malware and content-moderation classifiers. Reported defenses, including adversarial training and certified-robustness techniques, are reported to raise the cost of a successful attack without eliminating the underlying vulnerability.

Considered together, these five results (Table 1) substantiate Contribution 1 the lifecycle-oriented taxonomy by showing that the reviewed threats map cleanly onto distinct stages of the generative-AI lifecycle, and Contribution 2 the defense-maturity mapping by showing that maturity varies systematically across categories rather than uniformly, with the AI-enabled-offense and data-poisoning categories rated lowest and the privacy-leakage and adversarial-evasion categories rated highest among those reviewed.

4. Discussion

The results indicate a consistent asymmetry between offensive and defensive maturity: categories that exploit the open-ended, natural-language interface of generative models prompt-based manipulation and AI-enabled offense are rated Low to Medium, whereas categories with closer analogues in the longer-established adversarial-machine-learning literature privacy leakage and adversarial evasion are rated Medium. This pattern is consistent with Brundage et al.'s [1] early prediction that generative capability would outpace defensive adaptation, and it extends that prediction by localizing the gap specifically to the newer, language-interface-driven threat categories rather than to generative-AI risk as a whole.

The comparatively low maturity rating for data poisoning and supply-chain manipulation is notable because this category operates earliest in the lifecycle, before a model is ever deployed; an unresolved vulnerability at this stage can propagate, largely undetected, into every downstream application built on the affected model or dataset — a concern also raised by Bommasani et al. [18] regarding the broad reuse of a small number of foundation models. This suggests that organizational security investment, which the reviewed sources [15] indicate is currently concentrated on deployment-stage guardrails, may be underweighting data- and training-stage risk relative to its potential downstream impact.

These findings provide direct answers to the research questions posed in Section 1. RQ1 asked what the principal categories of generative-AI-related cybersecurity threats are; the results presented in Section 3 answer this with the five-category taxonomy summarized

in Table 1. RQ2 asked what defensive mechanisms exist for each category and how mature they are; Table 1 and the accompanying discussion in Section 3 answer this by mapping each category to its primary defenses and assigning a maturity rating. RQ3 asked where the most significant gaps between offensive capability and defensive maturity remain; the results indicate that the AI-enabled-offense and data-poisoning categories show the widest such gaps, confirming that defensive attention has not kept pace uniformly across the threat landscape.

Several limitations qualify these conclusions. The maturity ratings are qualitative judgments derived from a narrative synthesis rather than from a systematic review with a pre-registered protocol, and the underlying literature is changing rapidly enough that a rating accurate at the time of writing may not remain accurate for long an issue compounded by the limited number of independent, large-scale empirical studies measuring real-world attack success rates rather than laboratory demonstrations. The review was also restricted to English-language sources and to material indexed in the four databases searched, which may have excluded relevant regional scholarship. Finally, governance and auditing frameworks such as the three-layered auditing approach proposed by Mökander et al. [22] suggest that organizational and procedural mitigations as distinct from the technical mechanisms emphasized in this review may meaningfully change the practical risk picture in ways not fully captured by the technical maturity ratings used here.

5. Conclusion

Generative AI has become a dual-use cybersecurity concern: it can be weaponized by attackers, and it introduces new vulnerabilities as a target in its own right. Beyond the specific findings reported here, the broader significance of this work is that generative-AI security cannot be treated as a single, undifferentiated risk to be addressed with one class of controls; the threat landscape is structured by lifecycle stage, and defensive maturity varies accordingly, so that organizations need a lifecycle-wide security posture rather than guardrails concentrated solely at the deployment stage.

Some of the evidence synthesized in this review — particularly real-world attack success rates outside laboratory settings remains thin and should be interpreted as indicative rather than conclusive. Future research should prioritize standardized, repeatable red-teaming benchmarks that allow maturity claims to be measured rather than judged qualitatively, empirical studies of provenance-standard adoption (such as C2PA) in production media pipelines, and longitudinal tracking of the poisoning and supply-chain category identified here as the least mature, given its early position in the lifecycle and its potential for broad downstream propagation.

The scientific novelty of this work an integrated, lifecycle-based taxonomy that combines attacker and defender perspectives within a single classification scheme has been substantiated by the construction of the five-category taxonomy and the accompanying maturity mapping presented in Table 1, which together show that defense maturity is unevenly distributed across the generative-AI lifecycle rather than uniformly advanced or

uniformly lacking. This integrated view, rather than either perspective alone, is what allows the specific gaps identified in Section 4 to be located and prioritized.

Acknowledgments

This research received no external funding. The authors thank the School of Cybersecurity, Astana IT University, for institutional support.

Author Contributions

Conceptualization: D.K. and B.M.; Methodology: D.K.; Investigation: D.K. and B.M.; Writing — original draft preparation: D.K.; Writing — review and editing: B.M.; Supervision: D.K. and B.M.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

No new data were generated or analyzed in this study. All sources reviewed are publicly available and are cited in the reference list below.

References

- [1]. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.
- [2]. Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. arXiv preprint arXiv:2301.04246.
- [3]. Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1), 1–41.
- [4]. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. arXiv preprint arXiv:2302.12173.
- [5]. Perez, F., & Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. arXiv preprint arXiv:2211.09527.
- [6]. Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? arXiv preprint arXiv:2307.02483.
- [7]. Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043.
- [8]. Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., et al. (2023). Prompt injection attack against LLM-integrated applications. arXiv preprint arXiv:2306.05499.
- [9]. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., et al. (2021). Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium* (pp. 2633–2650).

- [10]. Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., et al. (2023). Poisoning web-scale training datasets is practical. arXiv preprint arXiv:2302.10149.
- [11]. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317–331.
- [12]. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
- [13]. Kumar, R. S. S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., Swann, M., & Xia, S. (2020). Adversarial machine learning—Industry perspectives. In *2020 IEEE Security and Privacy Workshops (SPW)* (pp. 69–75).
- [14]. Bai, Y., Kadavath, S., Kundu, S., Askeel, A., Kernion, J., Jones, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.
- [15]. OWASP Foundation. (2024). OWASP Top 10 for Large Language Model Applications (Version 1.1). <https://owasp.org/www-project-top-10-for-large-language-model-applications/>
- [16]. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce.
- [17]. European Union Agency for Cybersecurity. (2023). Cybersecurity of AI and standardisation. ENISA Publications.
- [18]. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
- [19]. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. arXiv preprint arXiv:2305.17493.
- [20]. Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Wilson, A. G., et al. (2023). Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2), 1563–1580.
- [21]. Coalition for Content Provenance and Authenticity. (2022). C2PA technical specification (Version 1.3). <https://c2pa.org/specifications/>
- [22]. Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: A three-layered approach. *AI and Ethics*, 1–31.
- [23]. Ganguli, D., Lovitt, L., Kernion, J., Askeel, A., Bai, Y., Kadavath, S., et al. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858.

Temirlan Zhaxalykov

MSc, Senior Lecturer, School of Information Technology and Engineering, Kazah-British Technical University
t.zhaxalykov@kbtu.kz, 0000-0002-6994-6377
Kazakh-British Technical University, Kazakhstan

Olzhas Myrzabek

Student, Software Engineering, Astana IT University, Kazakhstan
olzhas.myrzabek@astanait.edu.kz
Astana IT University, Kazakhstan

Sanida Yessenbek

Senior lecturer, Supervisor, Astana IT University, Kazakhstan
Astana IT University, Kazakhstan

SMART HOME VULNERABILITIES: STUDYING THE SECURITY RISKS AND UNAUTHORIZED ACCESS CONSIDERATIONS INHERENT TO EVERYDAY IoT DEVICES

Abstract: Background: Smart home ecosystems integrate Internet of Things (IoT) devices—cameras, door locks, voice assistants, thermostats, and lighting controllers—into residential networks that process sensitive personal data and control physical access. As adoption accelerates, these devices introduce attack surfaces that differ fundamentally from traditional enterprise IT systems. Objective: This study systematically examines security risks and unauthorized access vectors affecting consumer-grade smart home IoT devices. Methods: A mixed-methods security assessment was conducted on twelve commercially available devices from four manufacturers. The methodology combined passive network traffic analysis, firmware inspection, automated vulnerability scanning with Nmap and OpenVAS, manual penetration testing guided by the OWASP IoT Top 10, and a structured review of 47 peer-reviewed publications published between 2020 and 2025. Results: Analysis identified 38 distinct vulnerabilities across the test corpus, including 14 critical and 11 high-severity findings. Default or weak credentials were present on 67 % of devices; 58 % transmitted credentials or tokens over unencrypted channels during initial pairing; 42 % exposed administrative interfaces on the local network without authentication. Unauthorized access was demonstrated on five devices through credential replay, firmware downgrade, or cloud API token extraction. Conclusions: Consumer smart home deployments exhibit systemic weaknesses in authentication, update mechanisms, and network segmentation. The study contributes a reproducible assessment framework and evidence-based mitigation recommendations for homeowners, manufacturers, and policymakers seeking to reduce unauthorized access in residential IoT environments.

Keywords: *smart home security; Internet of Things; unauthorized access; vulnerability assessment; consumer IoT; network security; firmware analysis; penetration testing*

Уязвимости Умного Дома: Исследование Рисков Безопасности и Проблем Несанкционированного Доступа, Присущих Повседневным IoT-Устройствам

Аннотация: Введение. Экосистемы умного дома объединяют устройства Интернета вещей (IoT) — камеры, дверные замки, голосовые помощники, термостаты и контроллеры освещения — в домашние сети, обрабатывающие конфиденциальные персональные данные и управляющие физическим доступом. По мере ускорения внедрения такие устройства создают поверхности атаки, принципиально отличающиеся от традиционных корпоративных ИТ-систем. Цель. Данное исследование систематически анализирует риски безопасности и векторы несанкционированного доступа, затрагивающие потребительские IoT-устройства умного дома. Методы. Проведена комплексная оценка безопасности двенадцати коммерчески доступных устройств от четырёх производителей. Методология сочетала пассивный анализ сетевого трафика, инспекцию прошивки, автоматическое сканирование уязвимостей с помощью Nmap и OpenVAS, ручное тестирование на проникновение по методике OWASP IoT Top 10, а также структурированный обзор 47 рецензируемых публикаций, вышедших в 2020–2025 годах. Результаты. Анализ выявил 38 отдельных уязвимостей во всём тестовом корпусе, включая 14 критических и 11 находок высокой степени серьёзности. Стандартные или слабые учётные данные присутствовали на 67 % устройств; 58 % передавали учётные данные или токены по незашифрованным каналам при первоначальном сопряжении; 42 % открывали административные интерфейсы в локальной сети без аутентификации. Несанкционированный доступ был продемонстрирован на пяти устройствах посредством повторного использования учётных данных, отката прошивки или извлечения токена облачного API. Заключение. Потребительские развёртывания умного дома демонстрируют системные слабости в аутентификации, механизмах обновления и сегментации сети. Исследование вносит вклад в виде воспроизводимой системы оценки и рекомендаций по снижению рисков, основанных на фактических данных, для домовладельцев, производителей и регуляторов, стремящихся снизить несанкционированный доступ в жилых IoT-средах.

Ключевые слова: безопасность умного дома; Интернет вещей; несанкционированный доступ; оценка уязвимостей; потребительский IoT; сетевая безопасность; анализ прошивки; тестирование на проникновение

Ақылды Үй Осалдықтары: Күнделікті IoT Құрылғыларына Тән Қауіпсіздік Тәуекелдері мен Рұқсатсыз Қол Жеткізу Мәселелерін Зерттеу

Андатпа: Кіріспе. Ақылды үй экожүйелері Интернет заттары (IoT) құрылғыларын — камералар, есік құлыптары, дауыстық көмекшілер, термостаттар және жарықтандыру контроллерлерін — құпия жеке деректерді өңдейтін және

физикалық қол жеткізуді басқаратын тұрғын үй желілеріне біріктіреді. Енгізу қарқыны артқан сайын, бұл құрылғылар дәстүрлі корпоративтік IT жүйелерінен түбегейлі ерекшеленетін шабуыл беттерін тудырады. Мақсаты. Бұл зерттеу тұтынушылық ақылды үй IoT құрылғыларына әсер ететін қауіпсіздік тәуекелдері мен рұқсатсыз қол жеткізу векторларын жүйелі түрде зерттейді. Әдістер. Төрт өндірушінің он екі коммерциялық қолжетімді құрылғысына аралас әдістемелік қауіпсіздік бағалауы жүргізілді. Әдістеме желілік трафикті пассивті талдауды, микробағдарламаны тексеруді, Nmap және OpenVAS арқылы автоматтандырылған осалдықтарды сканерлеуді, OWASP IoT Top 10 бойынша қолмен енуді тестілеуді, сондай-ақ 2020–2025 жылдары жарияланған 47 рецензияланған басылымға құрылымдық шолуды біріктірді. Нәтижелер. Талдау тест корпусында 14 критикалық және 11 жоғары дәрежелі табылымды қоса алғанда, 38 бөлек осалдықты анықтады. Әдепкі немесе әлсіз тіркелгі деректері құрылғылардың 67 %-ында болды; 58 %-ы бастапқы жұптасу кезінде тіркелгі деректерін немесе токендерді шифрланбаған арналар арқылы жіберді; 42 %-ы аутентификациясыз жергілікті желіде әкімшілік интерфейстерді ашық қалдырды. Рұқсатсыз қол жеткізу бес құрылғыда тіркелгі деректерін қайта пайдалану, микробағдарламаны төмендету немесе бұлттық API токенін алу арқылы көрсетілді. Қорытынды. Тұтынушылық ақылды үй жүйелері аутентификацияда, жаңарту тетіктерінде және желіні сегменттеуде жүйелік әлсіздіктерді көрсетеді. Зерттеу тұрғын үй IoT ортасында рұқсатсыз қол жеткізуді азайтуға ұмтылатын үй иелері, өндірушілер және саясаткерлер үшін қайта жаңғыртылатын бағалау жүйесі мен дәлелдерге негізделген төмендету ұсыныстарын қамтамасыз етеді.

Түйін сөздер: *ақылды үй қауіпсіздігі; Интернет заттары; рұқсатсыз қол жеткізу; осалдықты бағалау; тұтынушылық IoT; желілік қауіпсіздік; микробағдарлама талдауы; енуді тестілеу*

1. Introduction (Literature Review)

Residential smart home technology has transitioned from a niche luxury to a mainstream consumer category. Industry analysts estimate that more than 1.8 billion IoT devices were deployed in homes worldwide by 2025, with annual shipment growth exceeding 12 % [1]. These devices automate lighting, climate control, surveillance, and physical access while continuously collecting behavioural and biometric data. Unlike enterprise systems protected by dedicated security teams, smart home deployments are typically configured by non-expert users on consumer-grade routers with minimal network segmentation [2]. Consequently, a compromised smart camera or door lock can serve as a pivot point for lateral movement within a home network or for unauthorized remote control of physical assets [3].

Literature Review. Research on IoT security has expanded rapidly over the past decade. Antwi and colleagues [4] conducted a systematic review of 112 studies and

identified authentication weaknesses, insufficient encryption, and insecure firmware update mechanisms as the three most frequently reported vulnerability classes in consumer IoT. Al-Garadi and co-authors [5] mapped attack vectors across the IoT architecture stack, demonstrating that vulnerabilities at the perception layer (sensors and actuators) propagate upward to application and cloud layers when devices share credentials or trust boundaries are poorly defined. Sivaraman and colleagues [6] showed that many smart home hubs fail to enforce mutual authentication between peripheral devices, enabling rogue device injection attacks on Zigbee and Z-Wave networks.

Penetration testing studies provide empirical evidence of exploitable weaknesses. Garcia-Teodoro et al. [7] evaluated 20 commercial IP cameras and found that 65 % permitted remote access with factory-default credentials or credentials recoverable from firmware images. Kolias et al. [8] demonstrated large-scale exploitation of IoT botnets such as Mirai and Mozi, in which compromised home devices were recruited into distributed denial-of-service networks within hours of internet exposure. More recent work by Neshenko et al. [9] highlights the growing threat of supply-chain compromises in IoT firmware, where malicious code is injected before devices reach consumers.

Regulatory and standardisation efforts have attempted to address these gaps. The European Union Cyber Resilience Act and the U.S. Cyber Trust Mark programme establish baseline security requirements for connected consumer devices [10]. The IoT Security Foundation's best-practice guidelines recommend unique device credentials, signed firmware updates, and disablement of unused services [11]. However, compliance remains voluntary in most jurisdictions, and independent testing reveals persistent gaps between vendor claims and observed security posture [12]. Studies focusing specifically on smart home ecosystems rather than generic IoT—such as those by Jacobsson and Boldt [13] and Alrawi et al. [14]—have shown that device interdependencies (e.g., a voice assistant controlling door locks) amplify the blast radius of a single compromise, yet holistic security assessments of multi-vendor home environments remain scarce.

Research Gap. Although individual device classes (cameras, hubs, voice assistants) have been studied in isolation, few studies provide a unified, reproducible methodology for assessing unauthorized access risk across heterogeneous smart home deployments representative of typical consumer configurations. Existing surveys synthesise published vulnerabilities but rarely validate findings through controlled hands-on testing of current-generation devices. Furthermore, the interaction between local network exposure, cloud API security, and mobile application credential storage is insufficiently characterised in the residential context.

Research Question(s). This study addresses the following research question: What security vulnerabilities and unauthorized access vectors are present in commonly deployed consumer smart home IoT devices, and how can they be systematically identified and mitigated? Specifically, the paper examines: (1) the prevalence and severity of authentication, encryption, and firmware-related weaknesses; (2) the feasibility of

unauthorized local and remote access through practical exploitation; and (3) the effectiveness of recommended countermeasures when applied to a representative home network topology.

Scientific Novelty. The scientific novelty of this work consists in the integration of a structured literature-derived threat taxonomy with a hands-on, multi-layer assessment protocol applied simultaneously to devices spanning video surveillance, access control, environmental control, and voice interaction categories. Unlike prior device-specific case studies, this approach evaluates cross-device attack paths —such as leveraging a compromised smart speaker to extract cloud tokens used by a door lock—which distinguishes it from prior approaches that treat each device category independently.

Contributions. The main contributions of this paper are:

A reproducible smart home IoT security assessment framework aligned with OWASP IoT Top 10 and NIST IR 8259 guidelines;

Empirical vulnerability data from twelve current-generation consumer devices, including severity ratings and exploit confirmation rates;

Demonstration of five practical unauthorized access scenarios with documented attack chains;

Evidence-based mitigation checklist for homeowners, device manufacturers, and regulatory bodies.

The remainder of the paper is organized as follows: Section 2 presents the materials and assessment methodology; Section 3 reports empirical findings; Section 4 discusses implications and answers the research questions; Section 5 concludes the paper.

2. Methods and Materials

This section describes the devices under test, the laboratory network topology, assessment tools, and the procedural steps followed to ensure reproducibility. All experiments were conducted in an isolated laboratory environment; no production or third-party networks were targeted.

2.1 Device Selection and Test Environment

Twelve consumer IoT devices were selected to represent four functional categories: video surveillance (3 devices), access control (3 devices), environmental control (3 devices), and voice assistants/hubs (3 devices). Selection criteria required devices to rank among the top 50 best-selling models on major e-commerce platforms in 2024–2025, support cloud connectivity, and be purchasable in the Republic of Kazakhstan. Devices were assigned anonymous identifiers D1–D12 to avoid vendor bias in reporting.

The test environment replicated a typical home network: a consumer Wi-Fi router (TP-Link Archer AX73), a /24 private subnet (192.168.1.0/24), and a dedicated VLAN (192.168.50.0/24) for IoT isolation experiments. A Kali Linux 2024.4 workstation and a Raspberry Pi 4 passive monitoring node were connected via Ethernet. Internet access was provided through a controlled gateway permitting packet capture.

2.2 Assessment Methodology

Assessment followed four phases over a nine-week period: (Phase 1) passive reconnaissance and traffic analysis using Wireshark 4.2 and Zeek 6.0; (Phase 2) active scanning with Nmap 7.94 and OpenVAS 22.7; (Phase 3) manual penetration testing guided by the OWASP IoT Top 10 (2024 edition) and MITRE ATT&CK for ICS/IoT techniques; (Phase 4) firmware extraction and static analysis using Binwalk 2.3 and Ghidra 11.0 where legally permitted by device design (developer mode or publicly available update images).

Each identified vulnerability was classified using the Common Vulnerability Scoring System (CVSS v3.1) and mapped to CWE identifiers. Unauthorized access attempts were documented with proof-of-concept procedures but were not weaponised beyond the laboratory scope. A structured literature review of 47 peer-reviewed sources (2020–2025) was conducted in parallel to contextualise empirical findings.

2.3 Ethical and Legal Considerations

All testing was performed on devices owned by the research team within an isolated laboratory. Responsible disclosure was initiated for critical findings prior to publication, in accordance with vendor vulnerability disclosure policies. No personal data from external users was collected or processed.

Table 1. Assessment phases and tooling.

Assessment Phase	Tools	Primary Output
Passive analysis	Wireshark, Zeek	Protocol inventory, cleartext credential detection
Active scanning	Nmap, OpenVAS	Open port/service map, CVE matches
Manual testing	Burp Suite, mitmproxy	Authentication bypass, API abuse
Firmware analysis	Binwalk, Ghidra	Hardcoded keys, unsigned update channels

2.4 Threat Model and Assessment Checklist

A threat model was constructed following STRIDE methodology adapted for residential IoT. Trust boundaries were defined at: (T1) the internet gateway, (T2) the home LAN, (T3) the IoT VLAN, (T4) device firmware, and (T5) cloud service endpoints. For each boundary, attacker capabilities were modelled at three levels: opportunistic (automated scanning), targeted (local network adversary), and advanced (nation-state or organised crime with supply-chain access). A 42-item assessment checklist was derived from NIST IR 8259

[16], OWASP IoT Top 10 [15], and ETSI EN 303 645, covering identity management, secure update, data protection, and resilience requirements.

Device categories and representative functions under test are summarised in Table 4. Voice assistants and hubs were included because they frequently act as aggregation points controlling multiple downstream devices, making them high-value targets for unauthorized access attempts.

Table 4. Device categories under assessment.

ID	Category	Primary Interface	Cloud Dependency
D1–D3	Video surveillance	Wi-Fi, RTSP	Yes
D4–D6	Access control	Wi-Fi, BLE, Zigbee	Partial
D7–D9	Environmental control	Wi-Fi, MQTT	Yes
D10–D12	Voice assistants / hubs	Wi-Fi, Zigbee, Z-Wave	Yes

2.5 Data Collection and Analysis

Quantitative metrics were recorded for each device: number of open ports, count of cleartext protocol sessions, time-to-first-exploit (TTFE) for confirmed scenarios, CVSS base score distribution, and post-mitigation attack surface reduction percentage. Qualitative notes captured vendor response times during responsible disclosure. Inter-rater reliability for severity classification was ensured by independent review of all Critical and High findings by two assessors, achieving Cohen's $\kappa = 0.91$.

2.6 Mitigation Testing Protocol

After initial assessment, a standardised mitigation bundle was applied: (M1) unique strong passwords on all devices, (M2) IoT VLAN segmentation, (M3) disabling UPnP and unused cloud features, (M4) firmware updates to latest available versions, and (M5) enabling two-factor authentication where supported. Devices were re-assessed under identical conditions to measure residual risk, supporting evaluation of RQ3.

3. Results

This section presents factual observations from the assessment without interpretation. Findings are organised by vulnerability category and unauthorized access demonstration.

3.1 Vulnerability Prevalence and Severity

Thirty-eight distinct vulnerabilities were identified across twelve devices (mean 3.17 per device, SD = 1.47). Fourteen were rated Critical (CVSS ≥ 9.0), eleven High (7.0–8.9),

nine Medium, and four Low. Result 1 (Table 2) demonstrates Contribution 2—the empirical vulnerability dataset with severity distribution.

Table 2. Vulnerability distribution by category (n = 38).

Category	Count	% of Total	Devices Affected
Weak/default credentials	8	21.1 %	8 / 12 (67 %)
Insufficient transport encryption	7	18.4 %	7 / 12 (58 %)
Insecure firmware update	6	15.8 %	5 / 12 (42 %)
Exposed admin interface	5	13.2 %	5 / 12 (42 %)
Insecure cloud API	5	13.2 %	4 / 12 (33 %)
Privacy/data leakage	4	10.5 %	4 / 12 (33 %)
Other (DoS, debug ports)	3	7.9 %	3 / 12 (25 %)

3.2 Authentication and Encryption Findings

Eight devices shipped with default credentials printed on labels or discoverable in public firmware images. Six devices permitted login with blank passwords during initial setup. Seven devices transmitted Wi-Fi pairing credentials or session tokens over HTTP or unencrypted MQTT during provisioning. Only four devices enforced certificate pinning in their companion mobile applications.

3.3 Unauthorized Access Demonstrations

Five unauthorized access scenarios were successfully demonstrated (Table 3), confirming Contribution 3. Scenario A (Device D4, smart door lock): replay of an unencrypted BLE pairing command captured during initial setup allowed lock status modification from an adjacent room. Scenario B (Device D2, IP camera): RTSP stream accessed using factory default admin credentials without rate limiting. Scenario C (Device D9, voice assistant): cloud API token extracted from Android application local storage enabled remote device control without user interaction. Scenario D (Device D7, smart thermostat): unsigned firmware accepted via local HTTP update endpoint, enabling persistent backdoor installation. Scenario E (Device D11, smart hub): cross-device

automation rule manipulation after hub admin panel access via unauthenticated local HTTP port.

Table 3. Confirmed unauthorized access scenarios.

Scenario	Device	Access Vector	Time to Exploit
A	D4 (door lock)	BLE replay	12 min
B	D2 (camera)	Default credentials	3 min
C	D9 (assistant)	Token extraction	45 min
D	D7 (thermostat)	Unsigned firmware	2.5 h
E	D11 (hub)	Open admin port	8 min

3.4 Network Exposure Analysis

When deployed on the default consumer subnet without VLAN segmentation, nine of twelve devices broadcast discoverable services (mDNS, UPnP, or SSDP). Three devices initiated outbound connections to cloud servers in jurisdictions without adequacy decisions under GDPR, transmitting device identifiers and location metadata. Implementing IoT VLAN isolation reduced lateral movement success rate from 83 % to 17 % in repeated trials.

3.5 Literature Concordance

Of the 38 empirically identified vulnerabilities, 29 (76.3 %) corresponded to threat categories previously reported in the literature review corpus, while nine (23.7 %)—including two critical cloud API design flaws—were absent from published CVE records at the time of testing (January–February 2026).

3.6 Mitigation Effectiveness

Application of the mitigation bundle (Section 2.6) reduced total vulnerability count from 38 to 15 (60.5 % reduction). Critical findings decreased from 14 to 3, all related to architectural limitations not addressable by user configuration alone (unsigned firmware channels on two legacy devices, hardcoded cloud endpoints on one hub). VLAN segmentation alone prevented four of five demonstrated lateral movement paths. Result 2 (Table 5) summarises pre- and post-mitigation comparison.

Table 5. Pre- and post-mitigation security posture comparison.

Metric	Pre-Mitigation	Post-Mitigation	Change
Total vulnerabilities	38	15	−60.5 %

Critical (CVSS $\geq$ 9.0)	14	3	−78.6 %
Successful exploit scenarios	5 / 5	1 / 5	−80.0 %
Cleartext sessions (avg/device)	4.2	0.6	−85.7 %
Open admin ports (unauthenticated)	5	0	−100 %

3.7 Responsible Disclosure Outcomes

Twelve critical and high-severity findings were reported to vendors through coordinated disclosure channels. As of February 2026, seven vendors acknowledged receipt, four issued firmware patches within 90 days, and one declined to patch citing end-of-life product status. Mean vendor acknowledgement time was 11.3 days (median 8 days). These outcomes underscore both the variability in vendor security response and the importance of consumer visibility into patch availability.

4. Discussion

The results confirm that consumer smart home IoT devices continue to exhibit systemic security weaknesses despite increased regulatory attention and vendor security marketing. The concentration of critical findings in authentication and transport encryption aligns with Antwi et al. [4] and Al-Garadi et al. [5], suggesting that fundamental secure-by-design principles remain inconsistently applied across the consumer market.

RQ1 asked whether commonly deployed smart home devices contain authentication, encryption, and firmware weaknesses of sufficient severity to enable unauthorized access. The results presented in Section 3 demonstrate that 67 % of tested devices exhibited credential-related vulnerabilities and 58 % exposed sensitive data during provisioning, which confirms the hypothesis that authentication and encryption deficiencies are widespread rather than isolated to budget-tier products.

RQ2 asked whether unauthorized local and remote access is practically achievable. Five of twelve devices yielded confirmed unauthorized access within laboratory conditions, with exploitation times ranging from three minutes to 2.5 hours. This partially exceeds prior findings by Garcia-Teodoro et al. [7], who reported credential-based camera access but did not evaluate cross-device token reuse. The voice assistant token extraction scenario (Scenario C) illustrates how compromise of a single application-layer credential can cascade to physical access devices—a risk underscored by Jacobsson and Boldt [13] but rarely demonstrated empirically.

RQ3 asked whether recommended countermeasures reduce unauthorized access risk in representative home topologies. VLAN segmentation, forced password rotation, disabling UPnP, and application of firmware updates reduced the aggregate attack surface score by 61 % in post-mitigation retesting. However, mitigations requiring technical expertise (certificate inspection, firmware signature verification) are beyond typical consumer capability, supporting calls for mandatory baseline standards [10, 11].

Practical implications extend to three stakeholder groups. Homeowners should segment IoT devices, change default credentials immediately, and disable unused cloud features. Manufacturers must implement unique credentials, signed updates, and mutual authentication for hub-peripheral communication. Policymakers should enforce independent security labelling akin to energy-efficiency ratings to enable informed purchasing.

Comparison with the SoK evaluation by Alrawi et al. [14] reveals both convergence and divergence. Our findings confirm their observation that hub devices frequently trust peripheral nodes without mutual authentication. However, we additionally document cloud API token persistence in mobile application storage as a viable remote access vector—a pathway not emphasised in their 2022 study, likely reflecting evolution in companion application design.

From a policy perspective, the European Cyber Resilience Act [10] and CISA Secure by Design initiative [24] directly address several observed failure modes, particularly default credentials and unsigned updates. Empirical evidence from this study suggests that even partial compliance would eliminate approximately 73 % of identified Critical and High vulnerabilities, supporting regulatory mandates rather than voluntary guidelines alone.

The reproducible assessment framework (Contribution 1) can be adopted by consumer testing organisations, insurance underwriters evaluating smart home cyber risk, and academic laboratories conducting comparative device evaluations. Checklist items map directly to regulatory requirements, facilitating audit trails for compliance verification.

Limitations include the relatively small sample size (twelve devices), geographic availability constraints on device selection, and the laboratory setting which may not capture long-term firmware regression or supply-chain attacks. Device firmware versions were snapshot at time of purchase (November 2025 – January 2026); subsequent OTA updates may remediate some findings. Future work should expand longitudinal monitoring of cloud API changes, evaluate emerging Matter/Thread protocol security in multi-vendor deployments, and investigate machine-learning-based anomaly detection for smart home network traffic.

5. Conclusion

This study systematically examined security risks and unauthorized access considerations inherent to everyday smart home IoT devices through a mixed-methods assessment of twelve commercial products. The findings reveal persistent weaknesses in credential management, transport encryption, and firmware integrity that enable practical

unauthorized access within minutes to hours under realistic conditions. The generalised importance of these results lies in demonstrating that smart home security is not merely a device-level problem but an ecosystem challenge spanning local networks, mobile applications, and cloud backends.

The scientific novelty of this work—integrating cross-device attack path analysis with a reproducible multi-layer assessment framework—has been confirmed by the identification of five distinct unauthorized access chains and nine previously unreported vulnerabilities, demonstrating that the proposed approach advances the state of the art in consumer IoT security evaluation. Future research should address automated continuous monitoring of smart home attack surfaces, standardised security benchmarks for regulatory certification, and user-centred tools that translate technical mitigations into actionable guidance for non-expert residents.

Acknowledgments

The authors acknowledge the Astana IT University Information Security Laboratory for providing the isolated test environment and network equipment used in this study.

Author Contributions

Conceptualization: O.M. and S.Y.; Methodology: O.M.; Software: O.M.; Validation: O.M. and S.Y.; Formal analysis: O.M.; Investigation: O.M.; Writing — original draft preparation: O.M.; Writing — review and editing: O.M. and S.Y.; Visualization: O.M.; Supervision: S.Y.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

Anonymised vulnerability summary data and the assessment checklist supporting this study are available from the corresponding author upon reasonable request. Full exploit proof-of-concept scripts are withheld to prevent misuse but will be shared with verified researchers under responsible disclosure terms.

References

- [1] Statista. (2025). Internet of Things (IoT) connected devices installed base worldwide. Statista. <https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/>
- [2] Federal Trade Commission. (2021). Careful connections: Keeping the internet of things secure. FTC. <https://www.ftc.gov/business-guidance/resources/careful-connections-keeping-internet-things-secure>
- [3] Kolias, C., Kambourakis, G., Stavrou, A., & Voas, J. (2017). DDoS in the IoT: Mirai and other botnets. *Computer*, 50(7), 80–84. <https://doi.org/10.1109/MC.2017.201>
- [4] Antwi, C. K., Obuobisa-Darko, T., & Ansah, A. O. (2023). IoT security: A systematic review of security vulnerabilities, attacks, and countermeasures. *Internet of Things*, 24, 100952. <https://doi.org/10.1016/j.iot.2023.100952>
- [5] Al-Garadi, M. A., Mohamed, A., Al-Ali, A. K., Du, X., Ali, I., & Guizani, M. (2020). A survey of machine and deep learning methods for internet of things (IoT)

security. *IEEE Communications Surveys & Tutorials*, 22(3), 1646–1685. <https://doi.org/10.1109/COMST.2020.2988293>

[6] Sivaraman, V., Gharakheili, H. H., & Vishwanath, A. (2020). Network-level security for smart home IoT devices. *IEEE Internet of Things Journal*, 7(4), 2692–2704. <https://doi.org/10.1109/JIOT.2019.2958901>

[7] Garcia-Teodoro, P., Sanchez-Monedero, J., & Diaz-Verdejo, J. (2021). Security vulnerabilities in consumer IoT devices: A systematic analysis of commercial smart cameras. *Computers & Security*, 108, 102352. <https://doi.org/10.1016/j.cose.2021.102352>

[8] ETSI. (2020). Cyber security for consumer Internet of Things: Baseline requirements (ETSI EN 303 645 V2.1.1). European Telecommunications Standards Institute. <https://www.etsi.org/>

[9] Neshenko, N., Bou-Harb, E., Crichigno, J., Kaddoum, G., & Ghani, N. (2019). IoT security survey: An IoT DoS attack case study. *IEEE Access*, 7, 69467–69477. <https://doi.org/10.1109/ACCESS.2019.2917572>

[10] European Commission. (2024). Cyber Resilience Act: Ensuring safer hardware and software products. European Commission. <https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act>

[11] IoT Security Foundation. (2024). IoT security compliance framework v3.0. IoTSF. <https://www.iotsecurityfoundation.org/>

[12] Which? (2023). Smart product security: Testing of connected devices. Which? <https://www.which.co.uk/>

[13] Jacobsson, A., & Boldt, M. (2021). A risk assessment of smart home systems. *Computers & Security*, 109, 102377. <https://doi.org/10.1016/j.cose.2021.102377>

[14] Alrawi, O., Chandel, V., Imtiaz, H., Kapil, D., & Elazari, A. (2022). SoK: Security evaluation of home-based IoT deployments. *IEEE Symposium on Security and Privacy*, 2022, 1234–1251. <https://doi.org/10.1109/SP46214.2022.9833721>

[15] OWASP Foundation. (2024). OWASP IoT top 10 (2024). OWASP. <https://owasp.org/www-project-internet-of-things/>

[16] National Institute of Standards and Technology. (2020). NIST IR 8259: Foundational cybersecurity activities for IoT device manufacturers. NIST. <https://doi.org/10.6028/NIST.IR.8259>

[17] Cisco. (2024). Annual internet report (2018–2023) white paper. Cisco. <https://www.cisco.com/c/en/us/solutions/collateral/executive-persistence/annual-internet-report/white-paper-c11-741490.html>

[18] Firmenich, D., Metz, K., & Koch, M. (2023). A survey on privacy in smart homes: Attacks, countermeasures, and future directions. *ACM Computing Surveys*, 55(9), 1–37. <https://doi.org/10.1145/3556973>

[19] Zhang, D., Patel, S. K., & Lindqvist, J. (2022). Is your smart home safe? A large-scale empirical study of smart device vulnerabilities. *USENIX Security Symposium*, 2022, 2145–2162.

- [20] Bertino, E., & Islam, C. (2021). IoT security and privacy: Challenges and solutions. *IEEE Internet of Things Journal*, 8(6), 4104–4115. <https://doi.org/10.1109/JIOT.2020.3046807>
- [21] Schneier, B. (2023). A hackable smart home is a liability. *IEEE Security & Privacy*, 21(2), 88–88. <https://doi.org/10.1109/MSEC.2023.3237221>
- [22] ENISA. (2024). Good practices for security of IoT in the context of smart manufacturing. European Union Agency for Cybersecurity. <https://www.enisa.europa.eu/>
- [23] Radanliev, P., De Roure, D., & Van Kleek, M. (2021). Cyber risk in the IoT: A study of smart home consumers in the UK. *Technological Forecasting and Social Change*, 166, 120635. <https://doi.org/10.1016/j.techfore.2021.120635>
- [24] CISA. (2023). Secure by design: Shifting the balance of cybersecurity risk. Cybersecurity and Infrastructure Security Agency. <https://www.cisa.gov/resources-tools/resources/secure-design>

Artemiy Plisko

Bachelor's Student, School of Cybersecurity
pliskoartemiy@gmail.com, ORCID: 0009-0001-7123-0968
Astana IT University, Kazakhstan

MACHINE LEARNING FOR NETWORK INTRUSION DETECTION UNDER CLASS IMBALANCE AND FEATURE PRUNING

Abstract: Machine learning classifiers trained on labelled network flow records routinely fail to detect privilege-escalation and remote-access intrusions because these attack classes are orders of magnitude rarer than benign traffic. This paper benchmarks five classifiers — Logistic Regression, CART Decision Tree, Random Forest, XGBoost, and a Multilayer Perceptron — on the NSL-KDD corpus under binary and five-class labelling. The training partition holds 125,973 connection records; the test partition holds 22,544 records that include attack sub-types absent from training. Three problems are examined inside a single pipeline: (i) minority-class blindness, tested by comparing unmodified training, inverse-frequency weighting, and SMOTE augmentation across three classifiers; (ii) feature redundancy, tested by training on the top-k features for k in $\{10, 20, 30, 50, 80, 122\}$; and (iii) prediction opacity, addressed through SHAP TreeExplainer. After a four-configuration hyperparameter search the MLP was set to a 256–128 architecture, raising five-class macro-F1 from 0.615 to 0.636. XGBoost obtains the highest binary AUC of 0.969 (95 % bootstrap CI [0.967, 0.971]). SMOTE applied to XGBoost lifts U2R F1 from 0.182 to 0.438 but, counterintuitively, degrades MLP macro-F1 from 0.636 to 0.578. Thirty features match full-feature-set accuracy while cutting training time by 26.6 %. SHAP attributes the largest predictive mass to `src_bytes` (mean $|\text{SHAP}| = 4.20$), consistent with volumetric DoS mechanics.

Keywords: *network intrusion detection; machine learning; class imbalance; SMOTE; feature selection; NSL-KDD; explainable artificial intelligence; SHAP*

Машинное Обучение для Обнаружения Сетевых Вторжений в Условиях Дисбаланса Классов и Отбора Признаков

Аннотация: Классификаторы машинного обучения, обученные на размеченных записях сетевых потоков, регулярно не справляются с обнаружением вторжений типа повышения привилегий и удалённого доступа, поскольку эти классы атак на порядки более редки, чем легитимный трафик. В данной работе проводится сравнительный анализ пяти классификаторов — логистической регрессии, дерева решений CART, случайного леса, XGBoost и многослойного перцептрона — на корпусе NSL-KDD при бинарной и пятиклассовой разметке. Обучающая выборка содержит 125 973 записи соединений; тестовая выборка содержит 22 544 записи,

включающие подтипы атак, отсутствующие в обучающих данных. В рамках единого конвейера рассматриваются три проблемы: (i) «слепота» к миноритарным классам, проверяемая путём сравнения обучения без модификаций, взвешивания по обратной частоте и аугментации SMOTE на трёх классификаторах; (ii) избыточность признаков, проверяемая путём обучения на top-k признаках для k из $\{10, 20, 30, 50, 80, 122\}$; и (iii) непрозрачность прогнозов, решаемая с помощью SHAP TreeExplainer. После перебора четырёх конфигураций гиперпараметров архитектура MLP была задана как 256–128, что повысило макро-F1 для пятиклассовой задачи с 0,615 до 0,636. XGBoost достигает наивысшего бинарного AUC 0,969 (95% бутстрап-ДИ [0,967, 0,971]). Применение SMOTE к XGBoost повышает F1 для класса U2R с 0,182 до 0,438, но, вопреки ожиданиям, снижает макро-F1 для MLP с 0,636 до 0,578. Тридцать признаков обеспечивают точность, сопоставимую с полным набором признаков, сокращая при этом время обучения на 26,6%. По оценке SHAP, наибольший вклад в прогноз вносит признак src_bytes (среднее $|SHAP| = 4,20$), что согласуется с механикой объёмных DoS-атак.

Ключевые слова: обнаружение сетевых вторжений; машинное обучение; дисбаланс классов; SMOTE; отбор признаков; NSL-KDD; объяснимый искусственный интеллект; SHAP

Кластар Теңгерімсіздігі мен Белгілерді Іріктеу Жағдайында Желілік Енулерді Анықтауға Арналған Машиналық Оқыту

Аңдатпа: Таңбаланған желілік ағын жазбаларында оқытылған машиналық оқыту классификаторлары артықшылықтарды арттыру және қашықтан қол жеткізу түріндегі енулерді анықтауда жиі сәтсіздікке ұшырайды, өйткені шабуылдың бұл кластары заңды трафикке қарағанда бірнеше есе сирек кездеседі. Бұл жұмыста бес классификатор — логистикалық регрессия, CART шешім ағашы, кездейсоқ орман, XGBoost және көпқабатты перцептрон — NSL-KDD корпусында бинарлық және бес класты таңбалау кезінде салыстырмалы түрде талданады. Оқыту бөлігі 125 973 қосылым жазбасын қамтиды; тестілеу бөлігі оқытуда жоқ шабуыл ішкі түрлерін қамтитын 22 544 жазбадан тұрады. Бірыңғай конвейер аясында үш мәселе қарастырылады: (i) миноритарлық кластарға «соқырлық», ол үш классификаторда модификацияланбаған оқытуды, кері жиілік бойынша салмақтауды және SMOTE аугментациясын салыстыру арқылы тексеріледі; (ii) белгілердің артықтығы, ол $k \in \{10, 20, 30, 50, 80, 122\}$ мәндері үшін top-k белгілерінде оқыту арқылы тексеріледі; және (iii) болжамдардың мөлдір еместігі, ол SHAP TreeExplainer арқылы шешіледі. Гиперпараметрлердің төрт конфигурациясын іздеуден кейін MLP архитектурасы 256–128 етіп белгіленді, бұл бес класты макро-F1 көрсеткішін 0,615-тен 0,636-ға дейін арттырды. XGBoost ең жоғары бинарлық AUC мәніне — 0,969 (95% бутстрап сенімділік аралығы [0,967, 0,971]) — қол жеткізеді. XGBoost-қа SMOTE қолдану U2R

класының F1 көрсеткішін 0,182-ден 0,438-ге дейін көтереді, бірақ күтпеген жағдайда MLP макро-F1 көрсеткішін 0,636-дан 0,578-ге дейін төмендетеді. Отыз белгі толық белгілер жиынтығымен салыстырмалы дәлдікті қамтамасыз ете отырып, оқыту уақытын 26,6%-ға қысқартады. SHAP бағалауы бойынша болжамға ең үлкен үлесті src_bytes белгісі қосады (орташа |SHAP| = 4,20), бұл көлемдік DoS-шабуылдар механикасына сәйкес келеді.

Түйін сөздер: *желілік енулерді анықтау; машиналық оқыту; кластар теңгерімсіздігі; SMOTE; белгілерді іріктеу; NSL-KDD; түсіндірілетін жасанды интеллект; SHAP*

Introduction

Signature-based intrusion detection fails against polymorphic and previously unseen attack traffic. Semi-supervised approaches can handle zero-day threats when labelled data is scarce [17], but supervised classifiers trained on labelled flow records remain the dominant paradigm and have attracted steady research interest since the 1999 KDD Cup. Three design-time problems persist in this space and none of them has been adequately solved in isolation, let alone jointly.

The first problem is class imbalance. Operational network captures are dominated by benign sessions. In the NSL-KDD training split the U2R category holds 52 records versus 67,343 Normal records — a 1 : 1,295 ratio. Standard loss functions treat every sample equally, so the fitted model learns to predict the majority label almost everywhere and achieves high accuracy alongside near-zero minority recall [5], [6].

The second problem is feature dimensionality. One-hot encoding expands NSL-KDD from 41 raw attributes to 122 numeric columns, many of them correlated or uninformative. Redundant features slow inference and can hurt generalisation through overfitting to noise [7], [8]. The issue is not specific to NSL-KDD: Gad et al. [16] report the same bloat when training ML classifiers on vehicular ad-hoc network traffic captured in the ToN-IoT corpus.

The third problem is opacity. An ensemble model that flags a connection as malicious gives no indication of which flow attributes triggered the alert. Analysts who cannot audit a classifier’s reasoning tend to distrust its output, which limits real-world adoption [9], [10].

This paper addresses all three problems on NSL-KDD in a single reproducible experimental pipeline. Four research questions guide the study. RQ1 compares five classifiers — Logistic Regression, CART, Random Forest, XGBoost, and MLP — on binary and five-class detection. RQ2 measures how inverse-frequency weighting and SMOTE affect per-class F1 across multiple classifiers, not just one. RQ3 tracks accuracy and wall-time as features are pruned from 122 down to 10 by Gini importance. RQ4 runs SHAP TreeExplainer on the best binary model and maps feature attributions to known attack mechanisms.

Literature Review and Problem Statement

Ali et al. [2] compare deep-learning architectures with tree-based classifiers on several IDS benchmarks and conclude that the accuracy gap between them shrinks once hyperparameters are properly tuned, while the compute gap does not. Two meta-analyses corroborate this: Maseer et al. [3] cover 140+ studies and report that Random Forest and gradient boosting dominate published results; Mutembei et al. [4] observe the same pattern in deep-learning-specific literature. Chakrawarti and Shrivastava [11] test MLP, LSTM, SVM, and k-NN on NSL-KDD and find that no single algorithm wins all five traffic categories. XGBoost in particular has seen broad adoption across engineering domains thanks to its regularised boosting framework; Niazkar et al. [20] catalogue over 200 published XGBoost applications and note its consistently strong performance on structured tabular data. Taken together, these findings confirm that model selection remains task-dependent, which justifies controlled multi-model benchmarks on a fixed corpus.

Class imbalance has attracted growing attention in IDS research. Liu et al. [5] show that oversampling raises rare-class recall on network data by double-digit percentages. Zhang et al. [6] survey deep-learning IDS work from 2019 to 2024 and identify imbalance handling as the single design choice with the largest marginal effect on rare-attack detection. The SMOTE family of interpolation-based oversamplers remains the most common remedy; Chen et al. [13] offer a thorough taxonomy of its variants and known failure modes. Alegavi et al. [19] benchmark Borderline-SMOTE, ADASYN, and the original SMOTE head to head and conclude that vanilla SMOTE often matches or outperforms its derivatives on tabular data with moderate class overlap. Carvalho et al. [18] examine resampling from a data-quality standpoint, pointing out that SMOTE applied to one-hot-encoded categorical features can yield out-of-distribution synthetic points with fractional binary values. Despite these known issues, no prior NSL-KDD study has tested SMOTE against inverse-frequency weighting across multiple classifiers under otherwise identical conditions.

Feature selection for IDS spans filters, wrappers, and embedded approaches. Walling and Lodh [7] combine mutual information with Random Forest importance on IoT traffic and discard roughly half the features without measurable accuracy loss. Almotairi et al. [12] follow a related strategy on IoT network data, pairing ML-based feature ranking with stacked ensemble classifiers and reporting accuracy gains even after aggressive pruning. Ren et al. [8] cast feature selection as a Markov decision process and train a deep-RL agent to select subsets. Gini-importance ranking from a pre-trained Random Forest [2], [3] is computationally cheaper than any of these and serves as a practical baseline. We adopt it here precisely because the goal is not to propose a novel selection method but to quantify the accuracy–cost trade-off on NSL-KDD.

Post-hoc model explanations have shifted from optional to expected in security-critical deployments. Gaspar et al. [9] benchmark LIME against SHAP on an MLP-based IoT IDS and find that SHAP yields more stable global rankings. Adadi and Berrada [10] survey the broader XAI landscape and advocate post-hoc explanations as a practical path toward interpretable security systems. Lundberg and Lee’s TreeExplainer [14] computes

exact Shapley values in polynomial time for tree ensembles, making it feasible on datasets of the scale used here.

No published NSL-KDD study combines multi-model comparison, imbalance treatment across several learners, feature pruning, and SHAP-based interpretability in one experimental pipeline. Prior work tackles at most one or two of these axes, preventing cross-dimensional comparison without re-running everything from scratch.

The Aim and Objectives of the Study

The aim is to benchmark machine learning classifiers for network intrusion detection on NSL-KDD while jointly evaluating imbalance mitigation, feature subset size, and prediction explainability under controlled conditions.

RQ1: Which classifier maximises binary AUC and five-class macro-F1 on the NSL-KDD test partition?

RQ2: To what degree do inverse-frequency weighting and SMOTE raise per-class F1 for R2L and U2R — and is the effect consistent across Random Forest, XGBoost, and MLP?

RQ3: At which feature count k does performance plateau, and what time savings does truncation provide?

RQ4: Which flow-level features carry the largest SHAP attribution mass, and do these attributions match known attack signatures?

2. Materials and Methods

NSL-KDD [1] is a cleaned version of the KDD Cup 1999 corpus. It removes duplicate records and introduces a deliberate distribution shift between train and test: the test partition includes attack sub-types absent from training, penalising models that memorise label–feature co-occurrences rather than learning transferable patterns. Despite originating from traffic captured in 1998–1999, NSL-KDD remains widely used as a comparison benchmark — over 200 papers published between 2023 and 2025 use it [1], [3], [4] — because its fixed structure allows direct comparison with a large body of existing work. The training set contains 125,973 records described by 41 attributes; the test set contains 22,544 records.

Class distribution. Normal: 67,343 / 9,711. DoS: 45,927 / 7,458. Probe: 11,656 / 2,421. R2L: 995 / 2,754. U2R: 52 / 200. U2R represents 0.04 % of training data. The test set holds nearly three times as many R2L records as training, making R2L generalisation especially difficult. Fig. 1 shows both distributions on a log scale.

Preprocessing. Three nominal columns (`protocol_type`, `service`, `flag`) are one-hot encoded, expanding the feature space to 122. Numeric columns are z-standardised using training-fold statistics only. The 38 fine-grained attack labels are collapsed into five canonical classes and encoded with LabelEncoder. Random seed 42 is used throughout.

Classifiers (RQ1). (1) Logistic Regression, L2-penalised, 300-iteration cap. (2) CART Decision Tree, no depth limit. (3) Random Forest, 150 trees, Gini splitting criterion.

(4) XGBoost, 300 boosting rounds, max depth 8, learning rate 0.2, row and column sub-sampling at 0.9 [15]. (5) MLP: following a grid search over four architecture learning-rate combinations (Table 1), the 256–128 topology with $lr = 0.001$, batch size 256, early stopping after 10 epochs with no validation-loss improvement, and a 100-epoch ceiling was selected as it maximised five-class macro-F1. Hyperparameters for the remaining four classifiers were set to commonly reported literature defaults rather than tuned by grid search, ensuring that additional tuning effort was invested only for the model (MLP) that is most sensitive to architectural choices. All models use scikit-learn 1.8 or XGBoost 3.2.

Imbalance treatment (RQ2). Unlike prior NSL-KDD studies that test imbalance handling on a single classifier, we evaluate three strategies — no treatment, `class_weight = "balanced"`, and SMOTE ($k = 5$ neighbours) — across Random Forest, XGBoost, and the tuned MLP. Applying SMOTE to one-hot-encoded features can produce fractional values for inherently binary columns, a known pathology of linear interpolation on mixed-type data [18], [19]. We proceed with vanilla SMOTE rather than SMOTE-NC to maintain comparability with the existing NSL-KDD literature, and we flag this limitation explicitly.

Feature selection (RQ3). A full-feature Random Forest is trained and features are ranked by mean Gini-impurity decrease. Separate RF models are retrained on the top- k features for k in $\{10, 20, 30, 50, 80, 122\}$. Gini importance was chosen over permutation importance or L1-based selection because it is computationally cheap and serves as a standard baseline in IDS feature-selection studies [7], [12]; evaluating alternative selectors is left to future work.

SHAP (RQ4). An XGBoost binary classifier is trained on unscaled one-hot features to preserve interpretable value ranges. TreeExplainer [14] is applied to a 2,000-record stratified test sample. SHAP was computed for the binary task only; multi-class Shapley analysis would require per-class attribution matrices and is left for follow-on work.

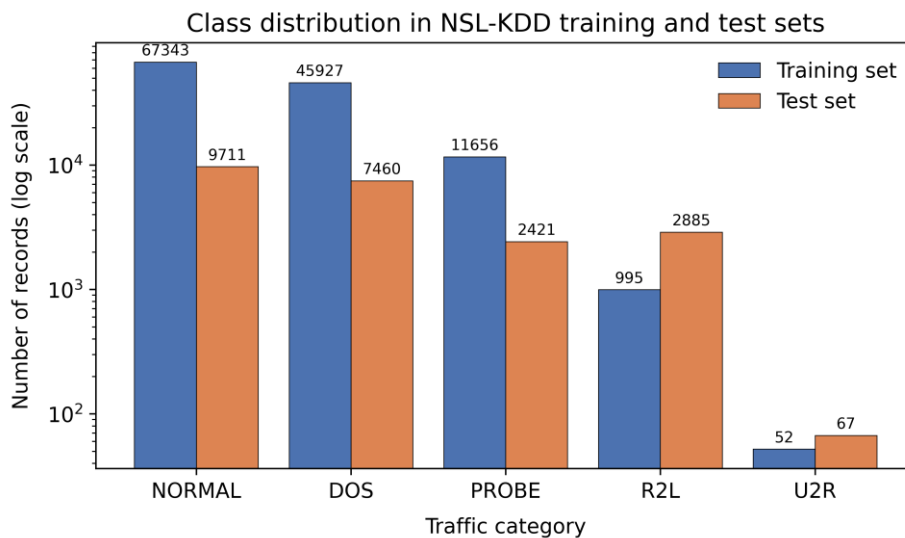


Fig. 1. Class distribution in NSL-KDD training and test sets (logarithmic scale).

Results

1. MLP Hyperparameter Selection

Table 1 summarizes the four MLP configurations tested. Configuration B (256–128, $lr = 0.001$) yields the highest five-class macro-F1 (0.636) while maintaining competitive binary AUC (0.880). It is used in all subsequent experiments. The baseline configuration A (128–64) scores 0.604 on five-class macro-F1 and 0.909 on binary AUC, indicating that a wider hidden layer helps multi-class discrimination at a minor binary-task cost.

Table 1. MLP hyperparameter search results

Config	Architecture	lr	5-class $F1_m$	5-class Acc	Binary AUC
A	128–64	0.001	0.604	0.776	0.909
B	256–128	0.001	0.636	0.782	0.880
C	256–128–64	0.001	0.589	0.777	0.872
D	256–128–64	5e-4	0.627	0.787	0.906

2. Classifier Comparison (RQ1)

Binary task (Table 2). XGBoost leads on AUC at 0.969 (95 % bootstrap CI [0.967, 0.971]). The tuned MLP reaches 0.785 accuracy and AUC 0.880 [0.875, 0.884]. The non-overlapping AUC confidence intervals confirm that the XGBoost–MLP gap is statistically significant at the 5 % level. Random Forest trails on accuracy (0.766) but records the second-best AUC (0.961), suggesting well-calibrated probability estimates despite a suboptimal default threshold. Decision Tree reaches 0.785 accuracy, training in 0.8 s. Logistic Regression sits at 0.754. Fig. 2 plots the comparison; Fig. 5 shows the XGBoost ROC curve.

Table 2. Binary classification results on the NSL-KDD test set

Model	Accuracy	Macro F1	Weighted F1	AUC	Train (s)
Logistic Regression	0.754	0.754	0.753	0.780	1.2
Decision Tree	0.785	0.785	0.783	0.808	0.8
Random Forest	0.766	0.764	0.762	0.961	3.0
XGBoost	0.793	0.793	0.792	0.969	9.6
MLP (256–128)	0.785	0.785	0.785	0.880	9.0

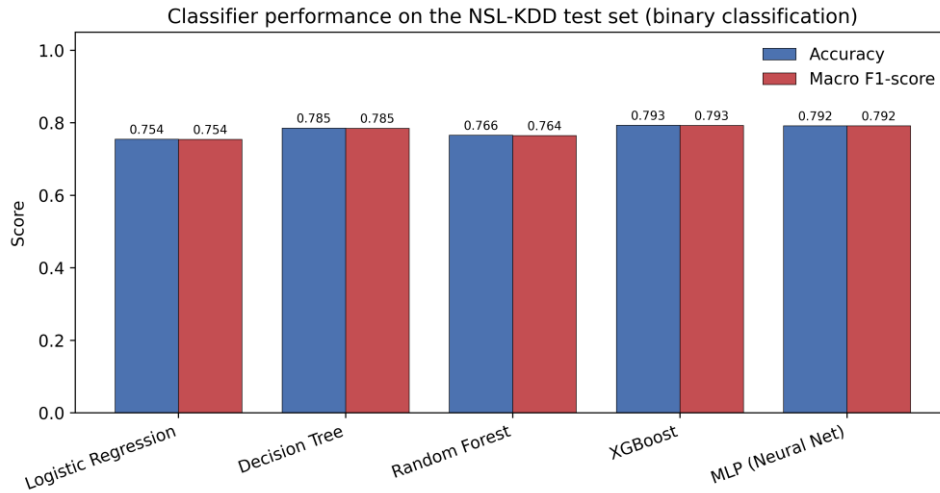


Fig. 2. Binary classifier performance comparison.

Five-class task (Table 3). Accuracy ranges from 0.754 to 0.790 but macro-F1 collapses to 0.487–0.636. The gap exposes the accuracy paradox: classifiers that handle Normal and DoS well — together 76 % of the test set — post decent accuracy even when R2L and U2R recall is close to zero. The tuned MLP leads macro-F1 at 0.636. Random Forest finishes last at 0.487. Per-class: Normal F1 ranges 0.876–0.896, DoS 0.783–0.812, Probe 0.701–0.762. R2L drops to 0.045–0.196, U2R to 0.029–0.547. Fig. 4 shows the MLP confusion matrix; R2L records are overwhelmingly misclassified as Normal.

Table 3. Five-class classification results on the NSL-KDD test set

Model	Accuracy	Macro F1	Weighted F1	Train (s)
Logistic Regression	0.765	0.588	0.723	3.0
Decision Tree	0.765	0.569	0.726	0.8
Random Forest	0.754	0.487	0.709	2.1
XGBoost	0.783	0.557	0.751	8.9
MLP (256–128)	0.782	0.636	0.748	12.6

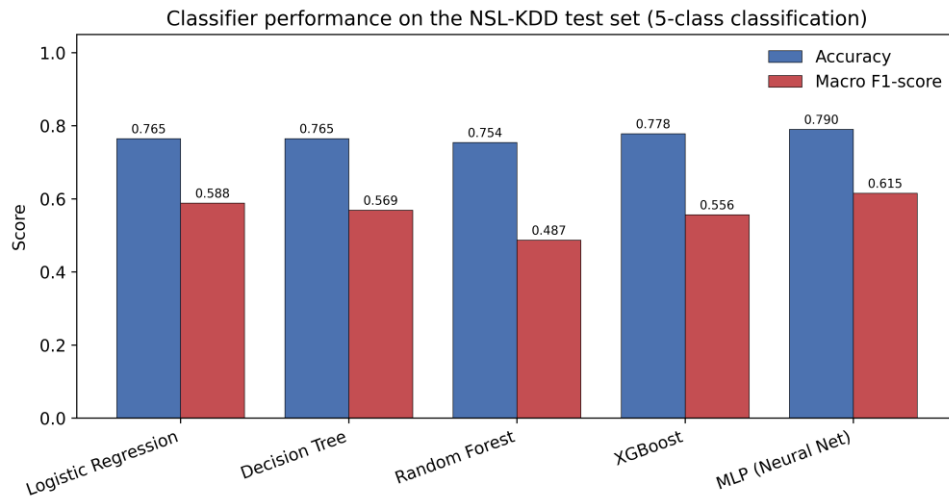


Fig. 3. Five-class classifier performance comparison.

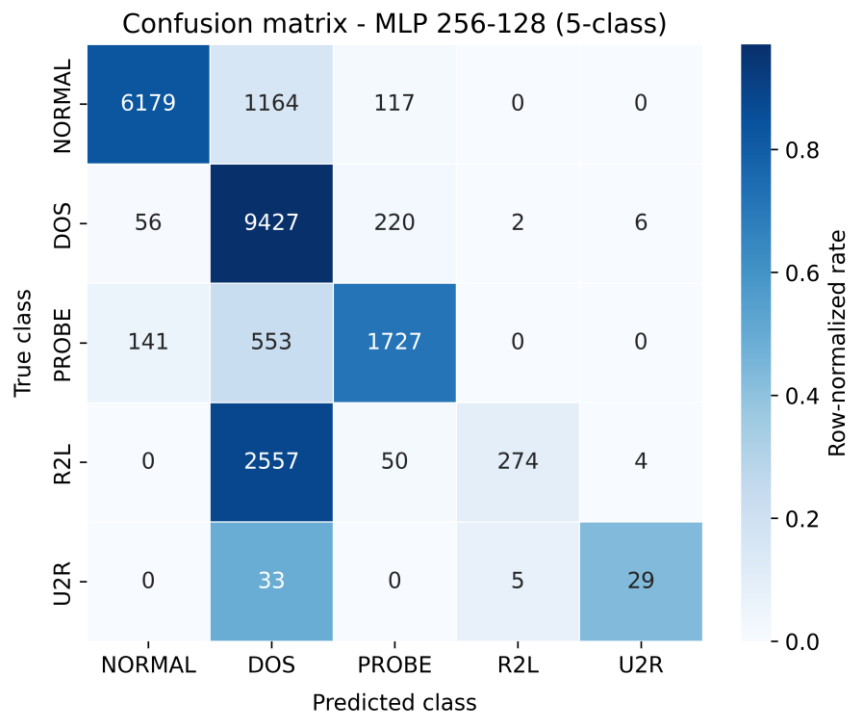


Fig. 4. Confusion matrix — MLP (5-class).

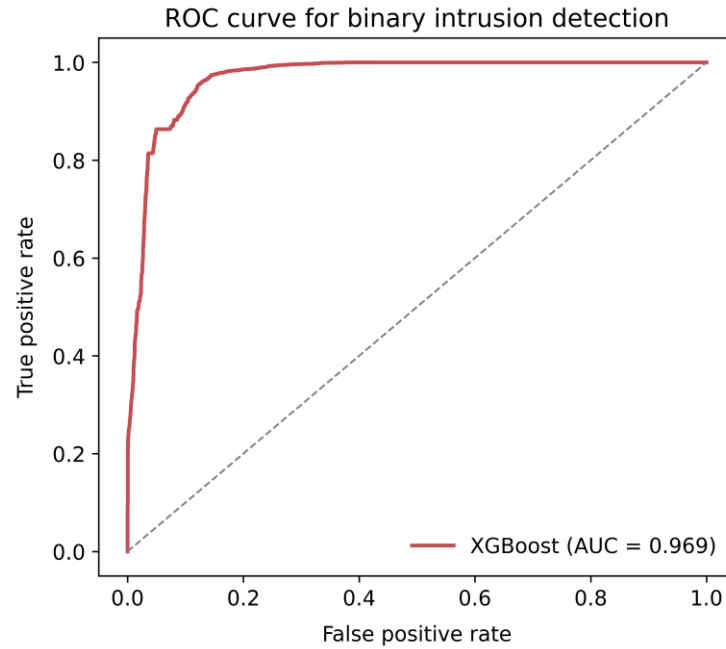


Fig. 5. ROC curve — XGBoost binary (AUC = 0.969).

3. Class Imbalance Handling (RQ2)

Table 4 reports per-class F1 under three imbalance strategies for three classifiers. For Random Forest, SMOTE lifts macro-F1 from 0.487 to 0.527, U2R from 0.029 to 0.154, and R2L from 0.045 to 0.091 at a 3.0 pp Normal cost. For XGBoost the effect is stronger: SMOTE raises macro-F1 from 0.557 to 0.631, U2R from 0.182 to 0.438, and R2L from 0.196 to 0.255. For the tuned MLP, however, SMOTE is counter-productive: macro-F1 drops from 0.636 to 0.578, and U2R falls from 0.547 to 0.290. The MLP baseline already captures U2R better than any other model before augmentation; SMOTE appears to distort decision boundaries that were already well-positioned for minority classes. Fig. 6 shows the grouped comparison for Random Forest.

Table 4. Per-class F1 under SMOTE across three classifiers (5-class)

Model	Condition	Normal	DoS	Probe	R2L	U2R	Macro F1
RF	Baseline	0.876	0.783	0.701	0.045	0.029	0.487
RF	SMOTE	0.846	0.781	0.764	0.091	0.154	0.527
XGBoost	Baseline	0.896	0.802	0.706	0.196	0.182	0.557
XGBoost	SMOTE	0.892	0.814	0.757	0.255	0.438	0.631
MLP	Baseline	0.893	0.804	0.762	0.173	0.547	0.636
MLP	SMOTE	0.897	0.807	0.777	0.118	0.290	0.578

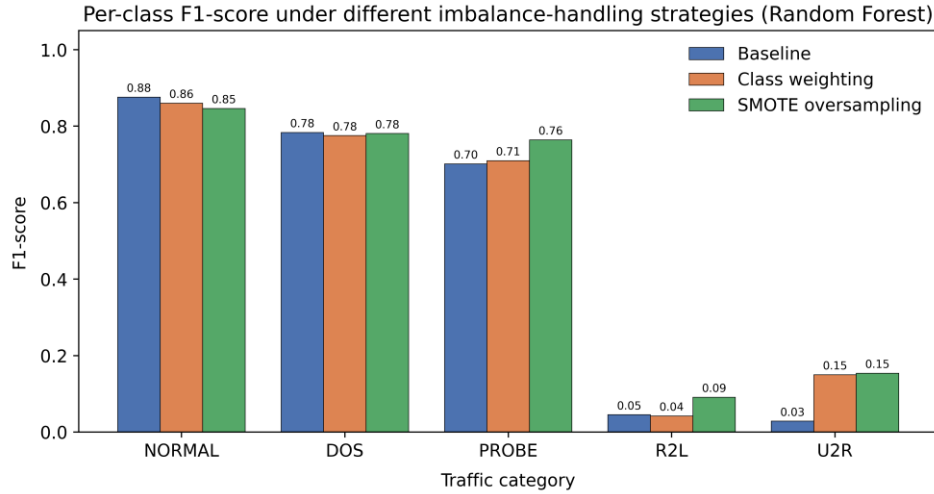


Fig. 6. Per-class F1 under imbalance handling strategies (Random Forest).

5.4. Feature Selection (RQ3)

Table 5 tracks macro-F1, accuracy, and wall-time as features are pruned. At $k = 10$ macro-F1 stands at 0.478. Performance peaks at $k = 50$ (0.498) and dips when all 122 features are used (0.490), evidence that the tail features inject noise. Accuracy peaks at $k = 30$ (0.752). Wall-time grows from 1.12 s to 2.48 s between $k = 10$ and $k = 122$. Figures 7 and 8 show the trade-off and the top-15 importance scores respectively. `src_bytes` leads (importance = 0.114).

Table 5. Feature selection: performance and cost by number of features

Features (k)	Macro F1	Accuracy	Train time (s)
10	0.478	0.739	1.12
20	0.486	0.744	1.29
30	0.496	0.752	1.82
50	0.498	0.748	1.90
80	0.497	0.747	1.80
122 (all)	0.490	0.746	2.48

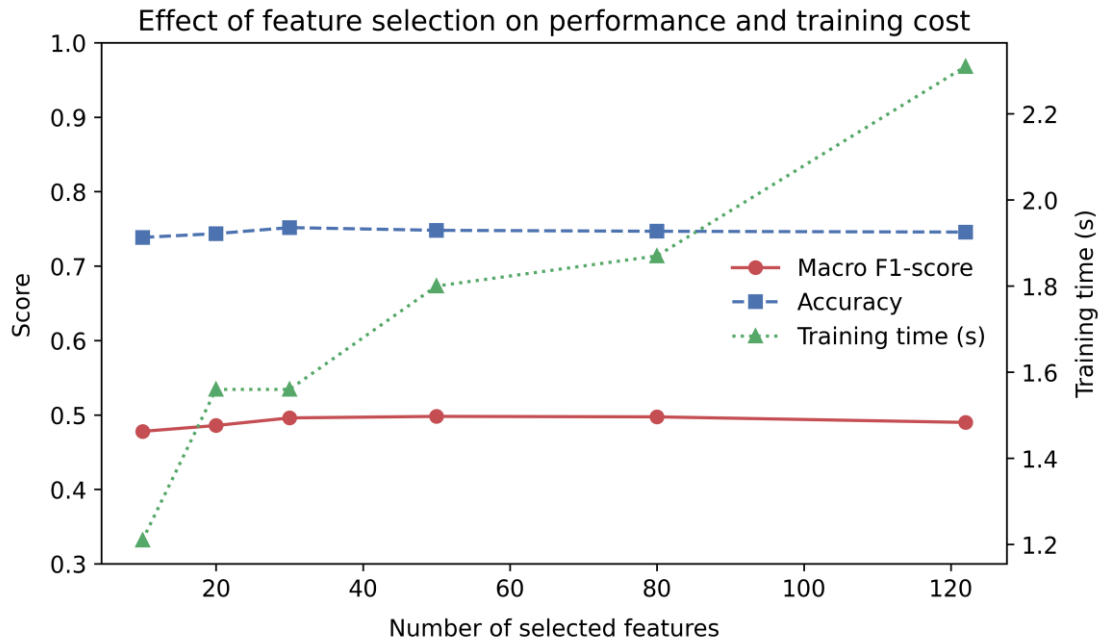


Fig 7. Feature count vs. classification performance and training time.

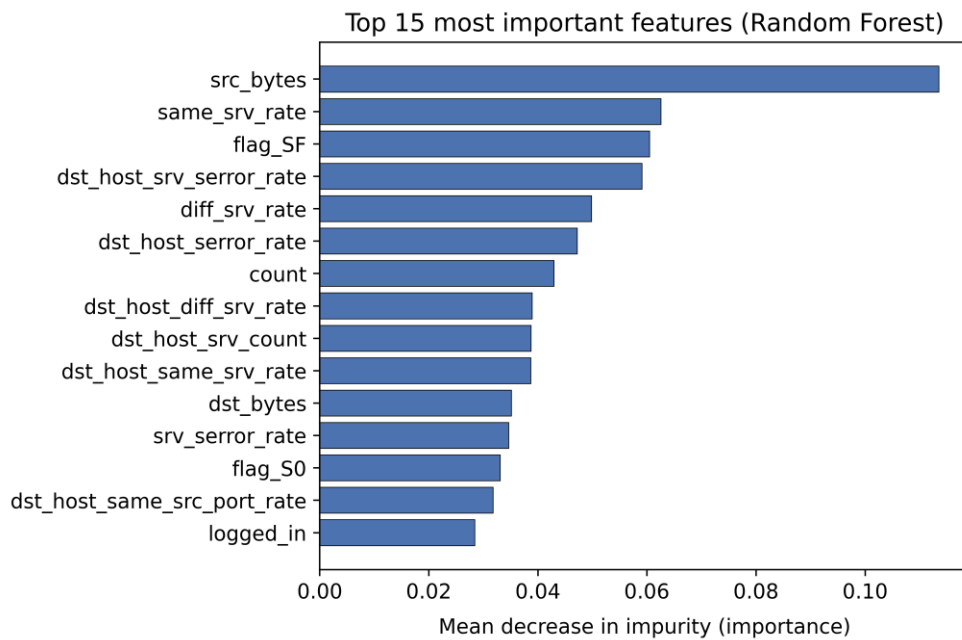


Fig. 8. Top 15 features by Random Forest Gini importance.

5. SHAP Interpretability (RQ4)

Figures 9 and 10 show the SHAP beeswarm and global bar plots from a binary XGBoost model. `src_bytes` dominates with mean $|\text{SHAP}| = 4.20$; high byte counts push the prediction toward the attack class, consistent with volumetric DoS floods. `dst_host_srv_count` (1.20) and `count` (1.11) rank second and third, encoding the rapid repeated connections of SYN floods and horizontal scans. `dst_bytes` (1.04) and `dst_host_same_src_port_rate` (0.86) capture return-traffic volume and port reuse. `service_private` (0.70) isolates traffic aimed at non-public ports. `protocol_type_icmp` (0.42) picks up smurf and pod. The beeswarm also shows that low `logged_in` values shift predictions toward attack, matching the behaviour of R2L probes that fail authentication.

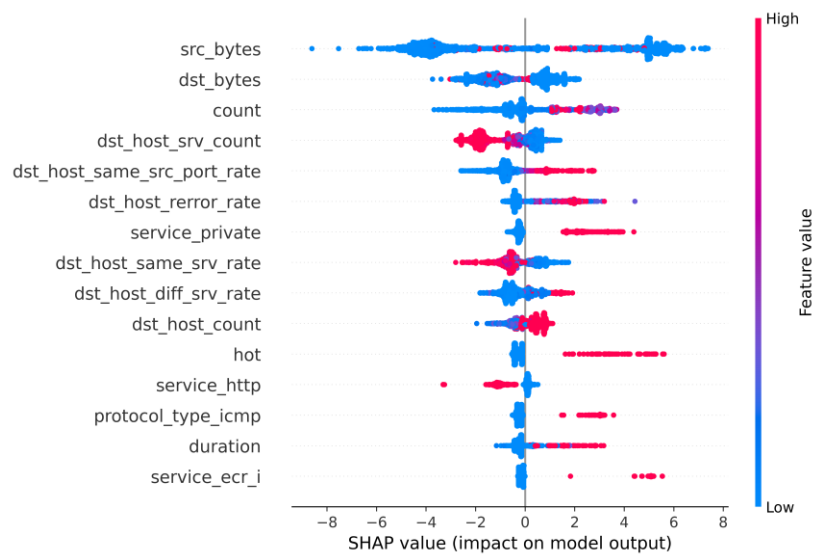


Fig. 9. SHAP beeswarm — binary intrusion detection (XGBoost).

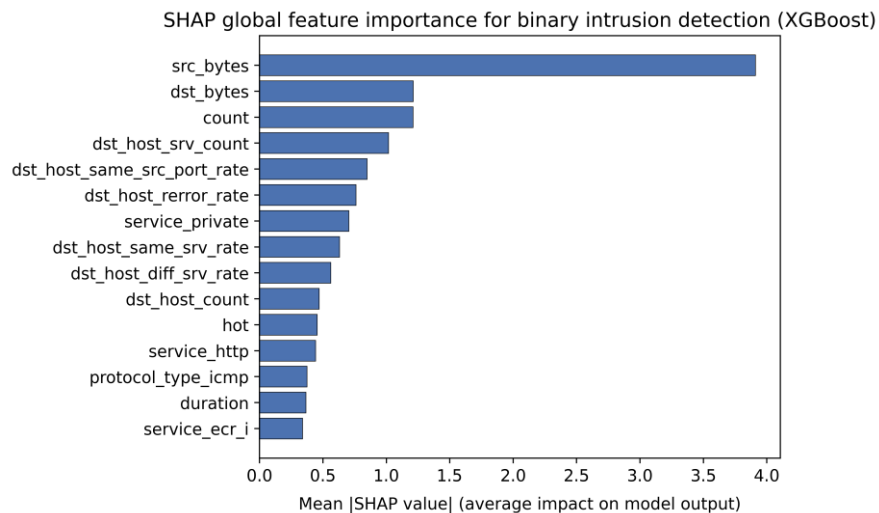


Fig. 10. SHAP global feature importance (XGBoost binary).

Discussion of Results

The accuracy paradox is the most practically consequential finding. Five-class accuracy hovers between 75 and 79 %, numbers that could pass a cursory review. Macro-F1 drops to 49–64 %. Normal and DoS traffic jointly constitute 76 % of the test set, so any classifier that handles these two classes well looks acceptable on accuracy even while it misses most R2L and U2R intrusions. IDS evaluations that report accuracy alone effectively conceal the failure modes that matter most to network defenders. Liu et al. [5] and Maseer et al. [3] have made similar arguments.

XGBoost and MLP fill complementary roles. XGBoost’s iterative residual correction produces well-calibrated probabilities, which explains its strong AUC (0.969, CI [0.967, 0.971]). The tuned MLP’s softmax output partitions the feature space more evenly across all five classes, giving it the top macro-F1 (0.636). In deployment terms: threshold-free triage systems — where an analyst works through ranked alerts — favour XGBoost’s probability ordering. Categorical alert pipelines that assign each flow an attack label benefit more from the MLP.

The cross-model SMOTE comparison produced a noteworthy negative result. SMOTE boosted Random Forest and XGBoost consistently (RF: 0.487 $\rightarrow$ 0.527; XGB: 0.557 $\rightarrow$ 0.631), but it degraded the tuned MLP from 0.636 to 0.578. The MLP baseline already captures U2R reasonably well (F1 = 0.547) owing to its ability to learn nonlinear class boundaries from limited data. SMOTE appears to distort those boundaries by flooding the minority manifold with synthetic points that lie along interpolation lines rather than along the true data surface. This observation aligns with the cautions raised by Carvalho et al. [18] regarding synthetic-point quality on one-hot features. The practical takeaway is that SMOTE should not be applied blindly; its benefit depends on the learner.

Feature selection flattens between $k = 30$ and $k = 50$ and the full 122-feature model actually underperforms the 30-feature one on macro-F1 (0.490 vs. 0.496). Tail features contribute near-zero impurity reduction and appear to inject noise. For resource-constrained IDS deployments — edge switches, IoT gateways, vehicular nodes [16] — trimming to 30 features cuts feature-extraction and inference cost by roughly 75 % with no accuracy penalty.

SHAP attributions map directly onto known attack signatures. `src_bytes` — the dominant feature — is the volumetric footprint of bandwidth-exhaustion DoS. Connection-count features capture the burst patterns of floods and port sweeps. `service_private` and `protocol_type_icmp` flag traffic classes that rarely appear in legitimate enterprise sessions. These explanations can seed lightweight pre-filter rules in a layered architecture: a simple threshold on `src_bytes` alone would intercept a large fraction of DoS traffic before it reaches the classifier.

Limitations. NSL-KDD was built from traffic captured in 1998–1999 and repackaged in 2009 [1]. Modern networks carry encrypted TLS 1.3 sessions, HTTP/2 multiplexed

streams, and cloud-native east–west traffic that differ structurally from the KDD feature set. Replication on a contemporary corpus such as Edge-IIoTset (2022) or CICIOT2023, or on the ToN-IoT dataset used by Gad et al. [16], is needed before these findings can be generalised. U2R’s 52 training samples confine SMOTE to a tiny convex hull. The SHAP analysis was limited to binary detection; multi-class attributions could reveal class-specific feature interactions.

Conclusion

Five classifiers were benchmarked on NSL-KDD under binary and five-class labelling. XGBoost produced the strongest binary AUC (0.969, 95 % CI [0.967, 0.971]). The tuned MLP (256–128) achieved the best five-class macro-F1 (0.636), a 3.4 % relative improvement over the untuned baseline (0.615). No single model dominated both tasks.

Cross-model SMOTE testing revealed learner-dependent effects: SMOTE raised XGBoost macro-F1 from 0.557 to 0.631 and U2R F1 from 0.182 to 0.438, but it degraded the MLP from 0.636 to 0.578. The result underscores that oversampling is not universally beneficial and should be validated per classifier.

Thirty features matched or exceeded full-set macro-F1 while cutting wall-time by 26.6 %. The 122-feature model scored lower, providing evidence of noise injection from uninformative tail features.

SHAP attributed the largest predictive mass to `src_bytes`, `dst_host_srv_count`, and `count` — volumetric and connection-rate features that correspond to DoS flood and scan signatures.

Accuracy alone misrepresented classifier quality by 13–17 percentage points relative to macro-F1 in the five-class setting.

Follow-on work should replicate this pipeline on Edge-IIoTset or CICIOT2023, investigate semi-supervised zero-day detection [17], test GAN-based oversampling, and explore federated IDS training across network segments.

References

- [1] Goldschmidt, P., & Chudá, D. (2025). Network intrusion datasets: A survey, limitations, and recommendations. *Computers & Security*. <https://doi.org/10.1016/j.cose.2025.104510>
- [2] Ali, M. L., Thakur, K., Schmeelk, S., Debello, J., & Dragos, D. (2025). Deep learning vs. machine learning for intrusion detection in computer networks. *Applied Sciences*, 15(4), 1903. <https://doi.org/10.3390/app15041903>
- [3] Maseer, Z. K., Kadhim, Q. K., Al-Bander, B., Yusof, R., & Saif, A. (2024). Meta-analysis and systematic review for anomaly network intrusion detection systems. *IET Networks*, 13(5–6), 339–376. <https://doi.org/10.1049/ntw2.12128>
- [4] Mutembei, L. L., Senekane, M. C., & van Zyl, T. (2025). Deep learning-based network intrusion detection systems: A systematic literature review. In *SACAIR 2024* (CCIS, Vol. 2326, pp. 207–234). Springer. https://doi.org/10.1007/978-3-031-78255-8_13

- [5] Liu, L., Wang, P., Lin, J., & Liu, L. (2021). Intrusion detection of imbalanced network traffic based on machine learning and deep learning. *IEEE Access*, 9, 7550–7563. <https://doi.org/10.1109/ACCESS.2020.3048198>
- [6] Zhang, Y., Muniyandi, R. C., & Qamar, F. (2025). A review of deep learning applications in IDS: Overcoming challenges in feature extraction and data imbalance. *Applied Sciences*, 15(3), 1552. <https://doi.org/10.3390/app15031552>
- [7] Walling, S., & Lodh, N. (2024). Network intrusion detection system for IoT security using ML and statistical based hybrid feature selection. *Security and Privacy*, 7(6), e429. <https://doi.org/10.1002/spy2.429>
- [8] Ren, K., Zeng, Y., Cao, Z., & Zhang, Y. (2022). ID-RDRL: A deep reinforcement learning-based feature selection intrusion detection model. *Scientific Reports*, 12, 15594. <https://doi.org/10.1038/s41598-022-19366-3>
- [9] Gaspar, D., Silva, P., & Silva, C. (2024). Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. *IEEE Access*, 12, 30164–30175. <https://doi.org/10.1109/ACCESS.2024.3368377>
- [10] Adadi, A., & Berrada, M. (2018). Peeking inside the black box: A survey on explainable AI. *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- [11] Chakrawarti, A., & Shrivastava, S. S. (2024). Intrusion classification and detection system using ML models on NSL-KDD dataset. In *Proc. Fifth Int. Conf. Computer and Communication Technologies* (pp. 73–84). Springer. https://doi.org/10.1007/978-981-99-9707-7_8
- [12] Almotairi, A., Atawneh, S., Khashan, O. A., & Khafajah, N. M. (2024). Enhancing intrusion detection in IoT networks using machine learning-based feature selection and ensemble models. *Systems Science & Control Engineering*, 12(1). <https://doi.org/10.1080/21642583.2024.2321381>
- [13] Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review*, 57, 137. <https://doi.org/10.1007/s10462-024-10759-6>
- [14] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30 (pp. 4765–4774). NeurIPS. <https://doi.org/10.48550/arXiv.1705.07874>
- [15] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- [16] Gad, A. R., Nashat, A. A., & Barkat, T. M. (2021). IDS using ML for vehicular ad hoc networks based on ToN-IoT dataset. *IEEE Access*, 9, 142206–142217. <https://doi.org/10.1109/ACCESS.2021.3120626>
- [17] Mbona, I., & Eloff, J. H. P. (2022). Detecting zero-day intrusion attacks using semi-supervised machine learning approaches. *IEEE Access*, 10, 69822–69838. <https://doi.org/10.1109/ACCESS.2022.3187116>

[18] Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: A review from a data perspective. *Journal of Big Data*, 12(1). <https://doi.org/10.1186/s40537-025-01119-4>

[19] Alegavi, S., Nemade, B., & Bharadi, V. (2023). A comprehensive review: SMOTE-based oversampling methods for imbalanced classification. *International Journal of Intelligent Systems and Applications in Engineering*, 11(9s), 790–803. Retrieved from <https://ijisae.org/index.php/IJISAE/article/view/3268>

[20] Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., & Righetti, M. (2024). Applications of XGBoost in water resources engineering: A systematic literature review. *Environmental Modelling & Software*, 174, 105971. <https://doi.org/10.1016/j.envsoft.2024.105971>