IVAN FRANKO NATIONAL UNIVERSITY OF LVIV
UNIVERSITY OF THE NATIONAL EDUCATION COMMISSION, POLAND
STATE UNIVERSITY OF INFORMATION AND COMMUNICATION TECHNOLOGIES
SCIENCE ENTREPRENEURSHIP TECHNOLOGY UNIVERSITY
OPOLE UNIVERSITY OF TECHNOLOGY, POLAND
ROWAN UNIVERSITY, USA
HALMSTAD UNIVERSITY, SWEDEN
NGO "CYBERSECURITY SPECIALISTS ASSOCIATION"
CYBER DEFENSE ACADEMY (CDA)



SECURITY OF MODERN INFORMATION AND COMMUNICATION SYSTEMS

PROCEEDINGS

of the International Scientific and Technical Conference

24–26 September 2026
Lviv – Krakow

UDC [003.26+004+519.816]:004.056:65(063)
SMICS: Security of Modern Information and Communication Systems: Proceedings of the International Scientific and Technical Conference, Lviv – Krakow, 24–26 September 2026. Lviv: Ivan Franko National University of Lviv, 2026. 242 pp.

This volume contains papers and abstracts submitted by participants in the International Scientific and Technical Conference “SMICS: Security of Modern Information and Communication Systems”. The conference aims to present recent advances and research findings in all aspects of cybersecurity and information protection.

EDITORIAL BOARD

VENHERSKYI P. – Head of the Department of Cybersecurity, Doct. Phys.-Math. Sci., Ivan Franko National University of Lviv, Ukraine;
GUTIK O. – Professor, Doct. Phys.-Math. Sci., Department of Cybersecurity Ivan Franko National University of Lviv, Ukraine;
TRUSHEVSKYY V. – Associate Professor, Cand. Phys.-Math. Sci., Department of Cybersecurity Ivan Franko National University of Lviv, Ukraine.

COMPILATION AND TYPESETTING

LEVCHENKO O. – Associate Professor, Cand. Tech. Sci., Department of Cybersecurity Ivan Franko National University of Lviv, Ukraine.

© 2026 Department of Cybersecurity
Ivan Franko National University of Lviv

ORGANIZERS

1. Ivan Franko National University of Lviv
2. University of the National Education Commission, Poland
3. State University of Information and Communication technologies
4. Science Entrepreneurship Technology University
5. Opole University of Technology, Poland
6. Rowan University, USA
7. Halmstad University, Sweden
8. NGO “Cybersecurity Specialists Association”
9. Cyber Defense Academy (CDA)

PARTNERS

















PROGRAMING COMMITTEE CO-CHAIR

BAK W. – Deputy Rector of the University of the National Education Commission, Krakow, dr hab. prof. (Poland)

GLADYSHEVSKII R. – Rector of Ivan Franko National University of Lviv, Academician of the National Academy of Sciences of Ukraine, Doctor of Chemical Sciences, Professor (Ukraine)

PROGRAM COMMITTEE

BUHRII O. – Vice-Rector for Research, Teaching, Social and Humanitarian Development, Ivan Franko National University of Lviv, Doctor of Physical and Mathematical Sciences, Professor (Ukraine)

DAVYDENKO A. – Leading Research Fellow, Pukhov Institute for Modelling in Energy Engineering of the National Academy of Sciences of Ukraine, Doctor of Technical Sciences, Professor (Ukraine)

EDISON PIGNATON DE FREITAS – Professor & Future Industry Research Program Leader Halmstad University (Sweden)

HNATYSHIN V. – Department of Computer Science Head, Full Professor, Rowan University (USA)

IVANCHENKO Y. – Director of the Educational-scientific Institute of Cyber Security and Information Protection, State University of Information and Communication Technologies, Doctor of Technical Sciences, Professor (Ukraine)

KLYMASH M. –Head of the Department of Information and Communication Technologies, Lviv Polytechnic National University, Doctor of Technical Sciences, Professor (Ukraine)

KLOTS Y. – Head of the Department of Cybersecurity, Khmelnytskyi National University, Candidate of Technical Sciences, Associate Professor

KORCHENKO O. – First Vice-Rector, State University of Information and Communication Technologies, Doctor of Technical Sciences, Corresponding Member of the National Academy of Sciences of Ukraine, Doctor of Technical Sciences, Professor (Ukraine)

MAHMOUD RAHAT – Associate Professor Halmstad University(Sweden)

NOVIKOV O. – Director of the ESI of Physics and Technology, Corresponding Member of NASU, Dr. Tech. Sci., Prof. Igor Sikorsky Kyiv Polytechnic Institute (Ukraine)

OPIRSKYY R. – Head of the Department of Information Technology Security, Lviv Polytechnic National University, Doctor of Technical Sciences, Professor (Ukraine)

PARKHUTS L. – Professor of Dept. of IT Security, Dr. Tech. Sci., Professor Lviv Polytechnic National University (Ukraine)

RZASA M. – Head of the Department of Cybersecurity, Full Professor, Opole University of Technology (Poland)

TKACH Y. – Head of the Department of Cybersecurity and Mathematical Modeling, National University "Chernihiv Polytechnic", PhD in Technical Sciences, Doctor of Pedagogical Sciences, Professor (Ukraine)

TORSTENSON O. – Programme Director, Master's Programme in Network Forensics (Sweden)

UDOVYK I. – Dean of the Faculty of Information Technologies, National Technical University "Dnipro Polytechnic", Cand. Tech. Sci., Professor (Ukraine)

VASUTA O. – Director of the Institute of Information Technology, University of the National Education Commission, Krakow, dr hab. prof. (Poland)

SCIENTIFIC SECRETARY

GUTIK O. – Ivan Franko National University of Lviv, Ukraine

ORGANIZING COMMITTEE CO-CHAIR

SEMENOV S. – Krakow University of the National Education Commission, Poland
VENHERSKYI P. – Ivan Franko National University of Lviv,Ukraine

MEMBERS

BOBERSKYY Y. – Bachelor of Science in Cybersecurity, Ivan Franko National University of Lviv, Ukraine
HRYTSYSHYN O. – Assistant, Dept. of Cybersecurity Ivan Franko National University of Lviv, Ukraine
KARPIUK R. – Senior Lecturer, Dept. of Cybersecurity Ivan Franko National University of Lviv, Ukraine
KOKOVSKA Y. – Associate Professor, Cand. Phys.-Math. Sci. Ivan Franko National University of Lviv, Ukraine
KROPYVA M. – Senior Lecturer, Dept. of Cybersecurity Ivan Franko National University of Lviv, Ukraine
KRZACZEK B. – Deputy Director of the Institute of Security and Information Technology for IT, PhD, University of the National Education Commission, Krakow, Poland
KUZBYT Y. – Assistant, Dept. of Cybersecurity Ivan Franko National University of Lviv, Ukraine
KYRYCHENKO N. – Engineer, Dept. of Cybersecurity Ivan Franko National University of Lviv, Ukraine
MELNYK I. – Associate Professor, Cand. Phys.-Math. Sci. Ivan Franko National University of Lviv, Ukraine
POPADIUK O. – Associate Professor, PhD, Ivan Franko National University of Lviv, Ukraine
PRYGARA M. – Associate Professor, Cand. Tech. Sci., Uzhhorod National University, Ukraine
SHCHERBYNA M. – Senior Lecturer, Co-Chair of Organizing Committee
Ivan Franko National University of Lviv, Ukraine
TERLETSKYY O. – Assistant, Dept. of Cybersecurity
Ivan Franko National University of Lviv, Ukraine
TOMKIEWICZ L. – Employee of the Institute of Security and Information Technology for IT University of the National Education Commission, Krakow, Poland
TRUSHEVSKYY V. – Associate Professor, Cand. Phys.-Math. Sci., Ivan Franko National University of Lviv, Ukraine
VLASOV V. – Associate Professor, Cand. Phys.-Math. Sci., Ivan Franko National University of Lviv, Ukraine
ZLATOUS S. – Assistant, Dept. of Cybersecurity Ivan Franko National University of Lviv, Ukraine

TABLE OF CONTENTS

TRACK 1. CYBER DEFENSE OF COMMUNICATION SYSTEMS AND CLOUD INFRASTRUCTURE, INCIDENT MANAGEMENT....10

Assessing the Ability of Large Language Models to Reproduce Expert Classification of User Identification Parameters by Anonymity Levels....11
Ivan Azarov, Oleksandr Korchenko, Anna Korchenko and Illia Azarov....11

Principles for SIEM-Based Information Security and Incident Management in IT Companies....16
Olexander Belej, Iryna Artyshchuk, Kostyantyn Kolesnyk....16

Empirical Validation of an Azure-Based Data Classification Pipeline for SOC 2 Type 2 Compliance: Benchmarking Eight Large Language Models....22
Oleh Deineka, Oleh Harasymchuk....22

Cybersecurity: several major challenges....27
Mikolaj Karpinski, Szymon Czerkies, Magdalena Jankowicz, Michal Skowronek and Mikolaj Stach....27

Set-Theoretic Approach to Automating Cyber-Resilience Assessment of Critical Infrastructure Facilities....31
Yuliia Khokhlachova, Kateryna Khokhlachova....31

Mathematical Method and Criteria for Assessing the Physical Security of Server Hardware in Cloud Services....37
Yevhenii Pedchenko, Yevheniia Ivanchenko, Ihor Ivanchenko, Maryna Pedchenko and Yurii Korovaichenko....37

Autonomous Defensive Agents for Continuous Detection of AI-Powered Attacks on Educational Systems....42
Stepan Prokipchyn....42

Cognitive Correlation of Cyber Threats across OSI Model Layers for Incident Response Support....48
Dmytro Prokopovych-Tkachenko....48

An Integrated Decision-Support Loop for Cyber-Incident Neutralization in Military Command and Control: From the Evaluation Image to an Admissible Response....53
Hennadii Rybachok....53

Cross-Day Validation of FPR-Constrained Operating Thresholds for XGBoost-Based Botnet Detection....59
Vladyslav Samoilenko and Sergii Gakhov....59

Acoustic Anomaly Detection for Indoor Intrusion Using Ray-Traced Room Impulse Responses....64
Oleksandr Terletskyy, Valeriy Trushevskyy, and Ostap Hrytsyshyn....64

Методика виявлення фішингових електронних повідомлень на основі гібридних алгоритмів машинного навчання....69
Ігор Аверічев, Петро Поночовний....69

TRACK 2. METHODS OF COUNTERING DISINFORMATION INFLUENCE AND SOCIAL ENGINEERING....73

Eduard B. Fleißner and his textbook “Handbuch der Kryptographie”....74
Oleg Gutik, Mykola Shcherbyna and Yurii Boberskyy....74

Predictive Cyber Threat Intelligence: Combining NLP-Based Entity Extraction with Graph Link Prediction....78
Yevheniia Ivanchenko, Iryna Lozova and Polina Halushka....78

Human Competencies as a Risk Mitigation Factor in Disinformation and Social Engineering Attacks....83
Anna Kiełbus....83

A mathematical model for assessing the cyber resilience of an information resource with regard to input data adequacy....88
Oleksandra Kovalchuk, Eugenia Ivanchenko....88

Velocity Features as the Structural Invariant of Card-Testing Attacks in Online Payment Systems....93
Yurii Kuzhii and Yurii Furgala....93

Graph Centrality and Clustering Metrics for Identifying Automated Accounts in Social Networks....99
Nataliia Kyrychenko....99

C-Classically Prime Subsemimodules and Their Application to Semigroup Action Key Exchange....105
Ivanna Melnyk and Andriy Andrushko....105

Blockchain-based data integrity verification for multi-sensor flood prediction systems....109
Orest Onyshchenko and Yaryna Kokovska....109

Attribution, liability, and emerging norms in AI-driven cyber operations....115
Myron Patsai and Danylo Shcherbyna....115

Methods for Optimizing Resource Allocation in Cybersecurity Systems....121
Olha Popadiuk....121

Investigation and Testing of a Data Loss Prevention Method for LLM-Based Systems....126
Liliia Pushkar and Valeriy Trushevskyy....126

Mechanism of Cognitive Warfare Impact and Potential Approaches to Effective Counteraction....131
Stanislav Shumlianskyi, Serhiy Kostyria and Volodymyr Budz....131

AI-Driven Cybersecurity: Challenges and Research Directions....138
Olga Torstensson and Dmytro Prokopovych-Tkachenko....138

Event-Driven Security Monitoring for Serverless Architectures: Behavioral Analysis of Function and Service Interactions....142
Sviatoslav Zlatous and Petro Venherskyi....142

TRACK 3. SECURITY OF IOT DEVICES, CYBER-PHYSICAL SYSTEMS SOFTWARE, TECHNICAL SOLUTIONS FOR INFORMATION SECURITY....147

Multi-level protection system and access rights separation in the mobile application....148
Liubomyr Botseniuk and Mykhailo Rizak....148

A Methodology for Building Reproducible Cybersecurity Datasets for Internet of Things Devices....150
Oleksandr Cherkaskyi....150

Block Data Encryption Using Polynomial Modified Fibonacci Matrices....155
Pavlo Grytsiuk and Yurii Hrytsiuk....155

On Modern Theoretical Aspects of the Security of Syber-Physical Systems....162
Oksana Kholyavka, Nataliia Buhrii[2]and Oleh Buhrii....162

Adaptive Detection of Goal Conflicts in Multi-Agent Smart Home Cyber-Physical Systems....165
Ihor Krestianinov....165

Simulation of Fault Injection Attacks on a Secure Bootloader for the RISC-V Soft-IP Core (MIV_RV32) in Renode Framework....171
Liliia Melnychuk and Mykola Shcherbyna....171

Integration of neural network detectors into visual odometry systems....178
Mykhailo Moroz....178

Modeling Cyber-Physical Consequences of Additive Attacks on the Control Channel of an Aerodynamic Levitation System....182
Oleksii Novikov, Mykola Ilin, Iryna Stopochkina and Volodymyr Duduladenko....182

Information security and monitoring of the functioning processes of local information and communication systems....188
Oksana Onyshchuk....188

mazewall: restricting application workers on Linux....193
Leanid Piliptsevich....193

Digital-Twin-Referenced Physical-Layer Anomaly Detection for a 1-Bit RIS in 6G Networks....199
Andriy Riy, Marian Kyryk, Serhii Zablotskyi and Ivan Demydov....199

Modeling Critical States In Cyber System Elements Based On Alloy....205
Hennadii Shapovalov and Anatoly Kazakov....205

GP4-WLTE 4G Controller Backdoor as an Example of Unauthorized Access Risks in Chinese Equipment....215
Mykola Shcherbyna and Petro Venherskyi....215

Integration of CGAN and Federated Learning to Protect Decentralized SDN Networks from Adaptive EW....223
Maksym Stoyko and *Mykola Gnidenko....223*

Algorithms for Accelerating Modular Operations in Cryptographic Systems....225
Iryna Vinkovska and Lidiia Tymoshenko....225

A Resilience-Loss Model for Cloud-Oriented IoT Platforms with Resource-Constrained Embedded Devices....229
Volodymyr Voitovych and Roman Lys....229

Комбінований підхід до виділення опорних точок у стеганографії зображень....234
Ірина Борисенко....234

Адаптивний алгоритм гамування графічної інформації....238
Андрій Садченко, Олег Кушніренко....238



TRACK 1. CYBER DEFENSE OF COMMUNICATION SYSTEMS AND CLOUD INFRASTRUCTURE, INCIDENT MANAGEMENT

Assessing the Ability of Large Language Models to Reproduce Expert Classification of User Identification Parameters by Anonymity Levels

Ivan Azarov[1], Oleksandr Korchenko[1], Anna Korchenko[2] and Illia Azarov[1]

[1] *State University of Information and Communication Technologies, Solomianska 7, 03110 Kyiv, Ukraine*
[2] *Dnipro University of Technology, Avenue 19, 49005 Dnipro, Ukraine,*

Abstract

In the current study, an analysis was conducted on the ability of large language models (LLMs) to perform expert assessment for creating a reference table of user anonymity levels, involving 16 parameters and 5 levels of anonymity. In this research, an original reference table was generated by the author, against which the responses of popular models were compared, such as: claude-opus-5, claude-fable-5, claude-sonnet-5, gpt-5.6-sol, gpt-5.6-terra, gpt-5.6-luna, gemini-3.7-flash, gemini-3.1-pro, glm-5.3, glm-5.3-flash, grok-4.6, deepseek-v4-pro, deepseek-v4-flash, kimi-k3, qwen3.8-max, Minimax M3.

The responses of the specified models were analyzed, their ability to classify network and browser parameters of a digital fingerprint was described, and the key advantages and disadvantages of their application for identifying anonymous users were determined. In total, 256 assessments were obtained and validated. The maximum match with the reference for an individual model was 43.8% (7 matches out of 16), while the overall match rate was only 22.3% (57 matches out of 256). Furthermore, three parameters related to network lists were not used by any of the AI models. It was established that the optimal criterion for identifying anonymous users is not relying on the evaluation results of artificial intelligence models, but rather using them exclusively as an auxiliary tool with mandatory verification by a subject-matter expert to ensure the cyber protection of critical infrastructure information resources.

Keywords

Anonymous user identification, cyber defense of critical infrastructure, fingerprinting, large language models, expert assessment

1. Introduction

In the modern world, there is an active development of digital technologies.

One of the most popular directions of technological development is large language models (LLMs), which contain systematized global knowledge and help people automate routine processes, have more time to explore new fields of activity, and bring ideas to life. These models process both legitimate and illegitimate user requests, which leads to the development of new attack vectors against critical computing infrastructure [1-3].

Attackers use tools to disguise their connections as those of legitimate users in order to be able to carry out attacks on computing resources.

To counter these attacks, cybersecurity specialists must investigate these trends and the capabilities of various language models to protect critical infrastructure. In scientific studies [4-9], existing cybersecurity services and modern techniques of active and passive device fingerprinting were analyzed. The current scientific work focuses on the use of modern AI models to establish parameter baselines by anonymity level for identifying anonymous users.

Corresponding authors
azarovphone@gmail.com (I. Azarov); o.korchenko@duikt.edu.ua (O. Korchenko); annakor@ukr.net (A. Korchenko)
0000-0002-2732-8861 (I. Azarov); 0000-0003-3376-0631 (O. Korchenko); 0000-0003-0016-1966 (A. Korchenko); 0000-0002-6163-3499 (I. Azarov)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Therefore, the aim of this study is to evaluate the ability of modern artificial intelligence models to reproduce human expertise in the task of classifying user identification parameters by anonymity levels.

2. Conducting the experiment

In this scientific research, 16 modern top models were used, such as: claude-opus-5, claude-fable-5, claude-sonnet-5, gpt-5.6-sol, gpt-5.6-terra, gpt-5.6-luna, gemini-3.7-flash, gemini-3.1-pro, glm-5.3, glm-5.3-flash, grok-4.6, deepseek-v4-pro, deepseek-v4-flash, kimi-k3, qwen3.8-max, Minimax M3, to form a reference base for identifying anonymous users based on 16 parameters, such as: IP in the firehol tor database (TOREXITS, TE), IP in the firehol anonymous users database (FIREHOLANONYMOUS, FA), IP in the firehol proxies database (FIREHOLPROXIES, FP), Usage of anonymous services (AUP), Service Worker (SW), IndexedDB (IDB), Quota Incognito mode (AMU), Cookies Storage (SCK), Local Storage (SL), Session Storage (SS), Connection anonymizers (CAE), Content blockers (CSB), Browser anonymizers (UBA), WebRTC leak (WLU), DNS leak (DNSAU), Timezone leak (TZUB) across 5 anonymity levels, such as: low (LAL), below average (UNMAL), average (MAL), above average (UPMAL), and high (HAL), to compare how much these results differ from the reference base formed as a result of human evaluation [10-12].

Each of the parameters triggers at exactly one of the five anonymity levels.

To preserve the integrity of the experiment, an identical prompt was formed for these models without access to the reference evaluations.

The proportion of exact parameter matches according to our human expert assessment and the mean absolute deviation were calculated.

Since there were 16 models in the test, a complete sample containing 256 evaluations was formed.

To evaluate the results, the proportion of exact parameter matches according to our human expert judgment and the mean absolute deviation were calculated.

Let $N = 16$ - be the total number of parameters, $L_i \in \{1, 2, 3, 4, 5\}$ be the reference anonymity level for the i-th parameter (where the first corresponds to LAL, and the fifth to HAL), and $\hat{L}_i$ be the level determined by the model.

The number of exact matches (*match*) determines the number of parameters where the model's level matched the reference level $(L_i = \hat{L}_i)$.. The match percentage ($match_{pct}$), which conceptually corresponds to the classical classification accuracy metric [13], is calculated using formula (1):

$$match_{pct} = \left(\frac{match}{N} \right) * 100\% \quad (1)$$

The result is rounded to one decimal place. This metric is classified as strict, because a shift to an adjacent level is considered an error. Mean Absolute Deviation (MAD), which characterizes the degree of the model's error and is the optimal metric for evaluating the average performance of algorithms on ordinal scales [14], is calculated using formula (2):

$$MAD = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{L}_i - L_i \right| \quad (2)$$

The calculation result is rounded to three decimal places. Encoding the ordinal scale with numbers using equal intervals is defined as a mathematical simplification. The ranking of models is

formed in descending order of the number of matches (*match*), and in case of a tie, in ascending order of the deviation indicator (*MAD*).

Further analysis is conducted at three levels: the overall model (a row of 16 scores), an individual parameter (16 model scores compared to the reference one), and the entire matrix.

Analysis in the context of individual parameters is considered mandatory, because the aggregated indicator masks errors with opposite directions and does not allow identifying specific points of divergence between the expert's judgment and the conclusions of the artificial intelligence.

The calculation results are presented in Table 1.

3. Research results

Table 1.
Comparative table of expert evaluation results by AI models and a human.

Evaluator	TE	FA	FP	AUP	SW	IDB	AMU	SCK	SL	SS	CAE	CSB	UBA	WLU	DNSAU	TZUB	Num. of matches	Matches, %	MAD
Our reference	MAL	UNMAL	UNMAL	HAL	LAL	UNMAL	MAL	UNMAL	UNMAL	UNMAL	HAL	MAL	HAL	HAL	MAL	UPMAL	-	-	-
gpt-5.6-sol	HAL	UPMAL	MAL	UPMAL	MAL	UNMAL	MAL	UNMAL	UNMAL	LAL	HAL	LAL	HAL	HAL	HAL	UNMAL	7	3.8	0.938
qwen3.8-max	HAL	HAL	HAL	HAL	MAL	UPMAL	MAL	UNMAL	UNMAL	UPMAL	HAL	UNMAL	HAL	HAL	HAL	MAL	7	43.8	1.125
grok-4.6	HAL	UPMAL	UPMAL	HAL	LAL	UNMAL	LAL	LAL	UNMAL	MAL	HAL	LAL	HAL	UPMAL	UPMAL	MAL	6	37.5	0.938
minimax-m3	HAL	HAL	MAL	UPMAL	MAL	UNMAL	UNMAL	UNMAL	UNMAL	UNMAL	UPMAL	MAL	UPMAL	HAL	HAL	MAL	6	37.5	0.938
kimi-k3	HAL	UPMAL	MAL	UPMAL	UNMAL	UNMAL	UNMAL	LAL	UNMAL	UNMAL	UPMAL	LAL	HAL	HAL	UPMAL	MAL	5	31.2	0.875
glm-5.3-flash	HAL	UPMAL	UPMAL	UPMAL	UNMAL	MAL	UNMAL	UNMAL	UNMAL	UNMAL	UPMAL	UNMAL	UPMAL	UPMAL	MAL	MAL	4	25.0	0.938
claude-opus-5	HAL	HAL	UPMAL	UPMAL	UNMAL	LAL	UNMAL	MAL	UNMAL	MAL	HAL	LAL	HAL	HAL	UPMAL	MAL	4	25.0	1.062
deepseek-v4-pro	HAL	HAL	UPMAL	UPMAL	LAL	UNMAL	LAL	UNMAL	MAL	MAL	UPMAL	LAL	UPMAL	HAL	UPMAL	MAL	4	25.0	1.125
gpt-5.6-luna	UPMAL	MAL	MAL	UPMAL	UNMAL	UNMAL	MAL	LAL	LAL	LAL	UPMAL	LAL	HAL	UPMAL	UPMAL	UNMAL	3	18.8	0.938
gemini-3.1-pro	HAL	HAL	UPMAL	HAL	UNMAL	MAL	UNMAL	MAL	MAL	MAL	UPMAL	LAL	UPMAL	HAL	HAL	UPMAL	3	18.8	1.188
claude-fable-5	HAL	HAL	UPMAL	MAL	UNMAL	UNMAL	UNMAL	UNMAL	LAL	LAL	UPMAL	LAL	UPMAL	UPMAL	UPMAL	MAL	2	12.5	1.25
claude-sonnet-5	HAL	HAL	UPMAL	UPMAL	UNMAL	LAL	UNMAL	LAL	LAL	LAL	UPMAL	LAL	HAL	HAL	HAL	MAL	2	12.5	1.25
gemini-3.7-flash	HAL	HAL	UPMAL	UPMAL	UNMAL	MAL	UNMAL	UNMAL	MAL	MAL	UPMAL	LAL	UPMAL	HAL	HAL	MAL	2	12.5	1.25
glm-5.3	HAL	UPMAL	UPMAL	UPMAL	UNMAL	MAL	UNMAL	MAL	MAL	MAL	UPMAL	LAL	UPMAL	HAL	UPMAL	MAL	1	6.2	1.188
gpt-5.6-terra	HAL	UPMAL	UPMAL	UPMAL	UNMAL	LAL	UNMAL	LAL	LAL	LAL	UPMAL	LAL	HAL	UPMAL	UPMAL	UNMAL	1	6.2	1.25
deepseek-v4-flash	HAL	HAL	UPMAL	UPMAL	UNMAL	MAL	UNMAL	MAL	MAL	MAL	UPMAL	LAL	UPMAL	UPMAL	UPMAL	MAL	0	0.0	1.312

The most pronounced systemic discrepancy is observed during the analysis of three parameters related to network lists.

The presence of an IP address in the database of Tor nodes or proxy servers is interpreted by the models as almost unequivocal proof of anonymity, which explains the generation of scores at the *HAL* and *UPMAL* levels.

From an expert's perspective, this feature is evaluated moderately, as address databases are subject to change over time, and the detection of such an address indicates solely the use of a

specific communication channel, rather than the spoofing or concealment of the device's functionality.

A mirror situation is observed when analyzing browser extensions and masking signs: the level of the *CSB* and *TZUB* parameters was rated below the benchmark by 15 out of 16 models, while the *AUP* and *AMU* parameters were underestimated by 13 models, and *CAE* by 12.

In this case, the author's criterion for detecting spoofing and concealment of functionality comes into play, which is not applied by generative models: for them, the use of a content blocker or a time zone discrepancy with the actual geolocation is classified as legitimate functionality [15-17].

Furthermore, it was established that the model class does not determine the final result: models from different developers are located at the top of the ranking, while certain flagship solutions showed only 1-2 exact matches.

4. Conclusions

Based on the results of the scientific research, it was established that in the task of classifying 16 user identification parameters across five levels of anonymity, modern large language models are not yet capable of fully replicating expert judgment. The best result was recorded for the gpt-5.6-sol and qwen3.8-max models, amounting to 7 matches out of 16 (43.8%), while kimi-k3 had 5 matches (31.2%). In total, only 57 out of 256 assessments matched the benchmark (22.3%), and the results of deepseek-v4-flash did not match a single parameter (0 out of 16, i.e., 0.0%).

It was determined that the root cause of the identified systemic discrepancies lies not in the models' level of awareness, but in the applied evaluation criteria: the author's expert approach is based on verifying whether a certain parameter indicates the application, spoofing, or concealment of the functionalities of the hardware-software complex, whereas artificial intelligence only weighs the overall reputation of the feature in the context of anonymity.

Thus, it has been proven that the optimal approach for identifying anonymous users is the use of artificial intelligence models as an auxiliary tool for the efficient collection and processing of information, with mandatory verification by a domain expert.

References

[1] Alladi, T., Chamola, V., & Zeadally, S. (2020). Industrial control systems: Cyberattack trends and countermeasures. Computer Communications, 155, 1-8. https://doi.org/10.1016/j.comcom.2020.03.007

[2] Skopik, F., & Pahi, T. (2020). Under false flag: Using technical artifacts for cyber attack attribution. Cybersecurity, 3(1). https://doi.org/10.1186/s42400-020-00048-4

[3] Prasad, N., Diro, A., Warren, M., & Fernando, M. (2025). A survey of cyber threat attribution: Challenges, techniques, and future directions. Computers & Security, 157, 104606. https://doi.org/10.1016/j.cose.2025.104606

[4] Laperdrix, P., Bielova, N., Baudry, B., & Avoine, G. (2020). Browser fingerprinting: A survey. ACM Transactions on the Web, 14(2), 1-33. https://doi.org/10.1145/3386040

[5] Li, S., & Cao, Y. (2020). Who touched my browser fingerprint? A large-scale measurement study and classification of fingerprint dynamics. In Proceedings of the ACM Internet Measurement Conference (pp. 370-385). https://doi.org/10.1145/3419394.3423614

[6] Kumar, V., & Paul, K. (2023). Device fingerprinting for cyber-physical systems: A survey. ACM Computing Surveys, 55(14s), 1-41. https://doi.org/10.1145/3584944

[7] Zhang, D., Zhang, J., Bu, Y., Chen, B., Sun, C., & Wang, T. (2022). A survey of browser fingerprint research and application. Wireless Communications and Mobile Computing, 2022(1), 3363335. https://doi.org/10.1155/2022/3363335

[8] Azarov, I., Korchenko, O., Korchenko, A., Ivanchenko, I., & Azarov, I. (2025). Techniques for identifying anonymous users in critical infrastructure. In Digital Transformation: Strengthening the Cybersecurity Capacities in the Modern World: abstracts of the International Scientific and Practical Conference, Krakow, 4-5 November 2025 (pp. 1-2). State University of Information and Communication Technologies. https://duikt.edu.ua/uploads/p_2779_38079537.pdf

[9] Korchenko, O., Korchenko, A., & Azarov, I. (2025). Analysis of modern solutions for the identification of anonymous users. In Applied Innovations in Information and Communication Technology, Lecture Notes in Networks and Systems, vol. 1338 (pp. 284-319). Springer. https://doi.org/10.1007/978-3-031-89296-7_15

[10] Artificial Analysis. (2026). Independent Analysis of AI Models & AI Hosting Providers. Retrieved September 1, 2026, from https://artificialanalysis.ai/

[11] Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text annotation tasks. Proceedings of the National Academy of Sciences, 120(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

[12] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Advances in Neural Information Processing Systems 36 (pp. 46595-46623). https://doi.org/10.52202/075280-2020

[13] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep feedforward networks. In Deep Learning (pp. 161-217). MIT Press.

[14] Esuli, A., Fabris, A., Moreo, A., & Sebastiani, F. (2023). Learning to quantify. Springer Nature. https://doi.org/10.1007/978-3-031-20467-8

[15] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? Computational Linguistics, 50(1), 237-291. https://doi.org/10.1162/coli_a_00502

[16] Jansen, R., Wails, R., & Johnson, A. (2024). Repositioning real-world website fingerprinting on Tor. In Proceedings of the 23rd Workshop on Privacy in the Electronic Society (pp. 124-140). https://doi.org/10.1145/3689943.3695047

[17] Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernandez, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., & Testoni, A. (2025). LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 238-255). https://doi.org/10.18653/v1/2025.acl-short.20

Principles for SIEM-Based Information Security and Incident Management in IT Companies

Olexander Belej[1], Iryna Artyshchuk[2], Kostyantyn Kolesnyk[1]

[1] *Lviv Polytechnic National University, 79000 Lviv, Ukraine*
[2] *University of the National Education Commission, Podchorążych 2, 30-084 Kraków, Poland*

Abstract

This paper presents a conceptual set of principles for deploying Security Information and Event Management (SIEM) in an IT company. It examines the challenges of integrating heterogeneous event sources, maintaining asset information, and limiting false-positive alerts. The proposed sequence connects asset-data updates, event normalization and classification, aggregation, and review of correlation rules with incident response. It provides implementation guidance; its effects on alert quality and response time have not yet been evaluated with operational data.

Keywords

security information and event management, incident response, event correlation, asset management, security monitoring

1. Introduction

Security Information and Event Management (SIEM) systems collect and analyze events from multiple sources to support threat detection, incident response, and regulatory compliance [1]. Remote work and cloud services increase the diversity of events, placing demands on data coverage and processing capacity.

Effective SIEM deployment requires understanding business processes, integrating heterogeneous sources, adapting detection rules, and ensuring access to qualified staff [2]. These demands are particularly relevant to organizations with limited resources. Cloud services can reduce local infrastructure needs but require assessment of data-sharing risks.

Prior studies describe SIEM limitations and deployment practices [2, 3] and an event-based approach for IoT environments [4]. NIST guidance addresses enterprise log collection and management [5], while Jaeger et al. study distributed normalization of high-volume events [6]. González-Granadillo et al. examine alert correlation [7], and Bhatt et al. describe SIEM operations in security operations centers [8]. Cotroneo et al. evaluate security-alert filtering [9]; Cinque et al. propose contextual filtering and prioritization for application logs [10]. These studies address individual tasks; the present work connects asset-data updates, event processing, and rule review in an IT-company deployment.

The aim is to systematize implementation principles connecting current asset information with event processing and the review of SIEM correlation rules.

Corresponding authors
oleksandr.i.belei@lpnu.ua (O. Belej); iryna.artyshchuk@uken.krakow.pl (I. Artyshchuk); kostyantyn.k.kolesnyk@lpnu.ua (K.Kolesnyk)
0000-0003-4150-7425 (O. Belej); 0000-0001-7287-8451 (I. Artyshchuk); 0000-0001-9396-595X (K. Kolesnyk)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Statement of the problem

Traditional SIEM deployment involves costs for licensing, infrastructure, integration, maintenance, and staff training. These demands can be especially difficult for organizations with limited resources. Effective operation also requires regular updates to detection rules and careful selection of threat intelligence sources.

The quality of SIEM monitoring depends on balancing threat detection against the number of alerts generated. Incomplete data and poorly configured rules may cause threats to be missed, while overly broad rules can produce false positives and increase analysts' workload. Organizations must therefore review data sources, refine detection rules, and adapt their SIEM processes as threats and infrastructure change. Figure 1 summarizes the main challenges associated with SIEM use.

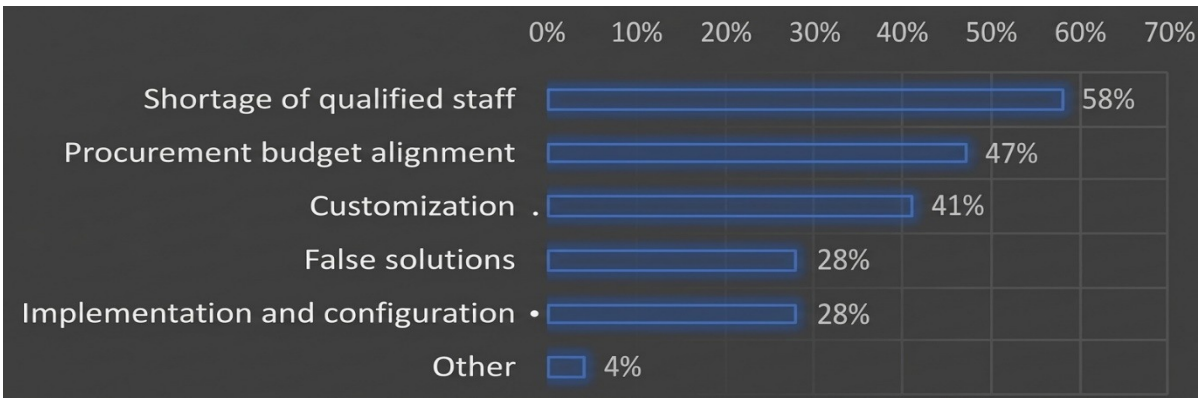


Figure 1: Main barriers to effective SIEM operation (adapted from [11]).

SIEM deployment is challenging because it requires the integration of heterogeneous data sources, sufficient processing capacity, and qualified personnel. Servers, applications, network devices, and security tools generate events in different formats; incomplete integration can create gaps in monitoring. Legacy systems may further complicate data collection and configuration.

Staffing is another constraint. In the 2022 SIEM Security Report, 41% of respondents identified a lack of skilled or trained staff as a hurdle to operating SIEM effectively [11]. Analysts must maintain rules and investigate alerts while processing many events. Log enrichment and contextual alerts can help prioritize investigations if data sources and rules are reviewed regularly.

3. Scientific novelty and proposed solution

Compared with a survey of SIEM capabilities [3] and a scenario-specific SIEM approach [4], this paper focuses on linking updates to monitored assets with event processing and rule review in an IT-company deployment. The scientific novelty lies in integrating dynamic asset-information updates, event normalization, aggregation, and periodic correlation-rule review into a unified asset-aware SIEM deployment sequence for IT companies. Access controls, antivirus tools, applications, and network equipment provide complementary incident data for this sequence.

The proposed steps are: inventory sources and assets; normalize source, actor, target, action, time, and outcome; classify events by process, entity, and attack type; aggregate related events while retaining logs; and review correlation rules as assets change. Logging and normalization

follow established guidance and research [5, 6]. This is an implementation sequence, not a tested algorithm.

Linking each event to a current asset identifier allows analysts to trace alerts to affected resources and prompts review of rules that depend on changed assets.

Operational evaluation should compare this sequence with an existing SIEM configuration using the same scenarios, measuring alert precision, missed incidents, and analyst review time. Evaluating the proposed approach using these measures is a goal of future research.

Before SIEM deployment, security functions are distributed across access-control, antivirus, monitoring, and intrusion-detection tools, which complicates coordinated event analysis. Figures 2 and 3 show how SIEM integrates these functions into a unified assessment and response workflow.

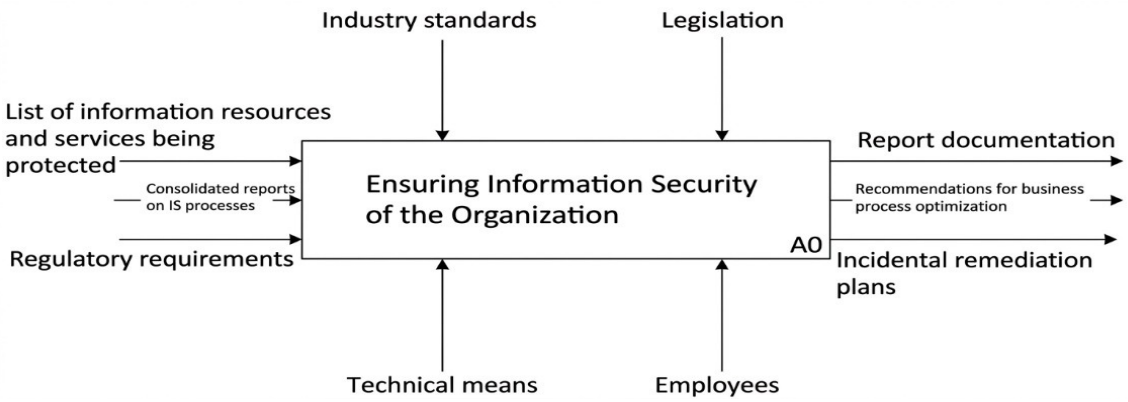


Figure 2: Information security process before SIEM integration.

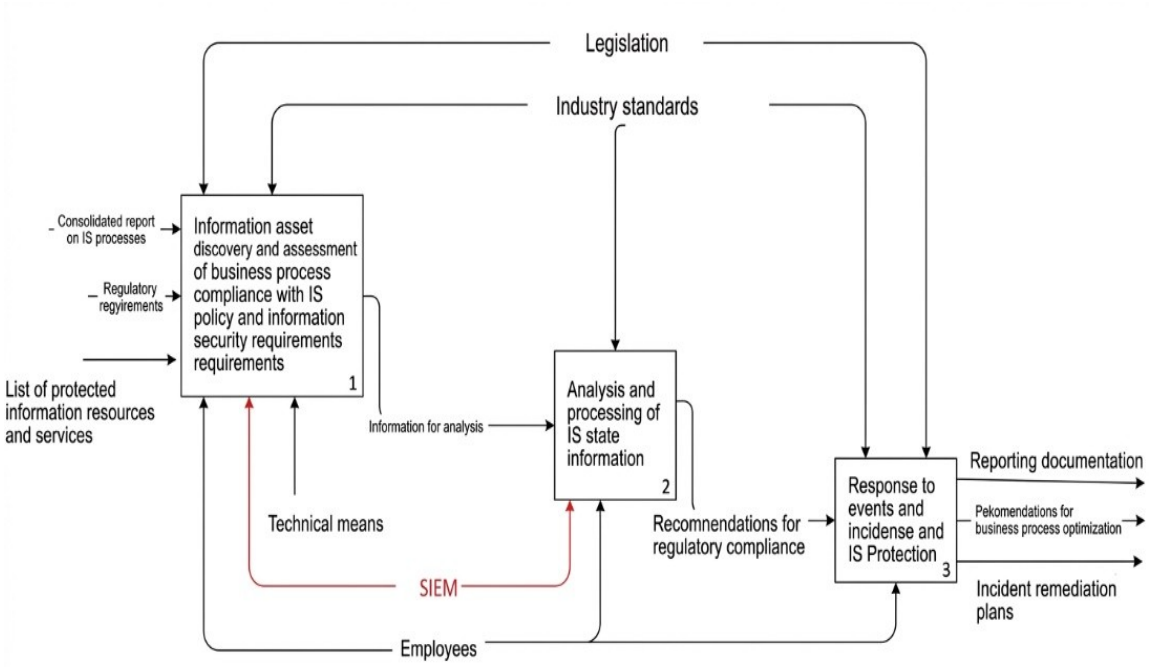


Figure 3: SIEM-supported information security assessment and incident response workflow.

As shown in Figure 3, SIEM links asset assessment and security-event analysis with incident response, regulatory compliance, and recommendations for improving business processes. The

principal advantage of SIEM is the centralized collection and correlation of security events, which improves threat detection, accelerates incident response, and reduces the need to monitor multiple security tools separately.

4. Discussion

Effective SIEM deployment depends on several organizational and technical conditions. Vendor-provided rules must be adapted to the organization's assets, data sources, infrastructure, and security policies. Asset information and detection rules must remain current as users, applications, devices, and network configurations change. Correlation rules, aggregation parameters, and alert thresholds should also be tested under realistic event loads to reduce false positives without concealing significant incidents. Consequently, SIEM effectiveness depends not only on automation but also on data quality, appropriate configuration, system integration, and continuous expert supervision.

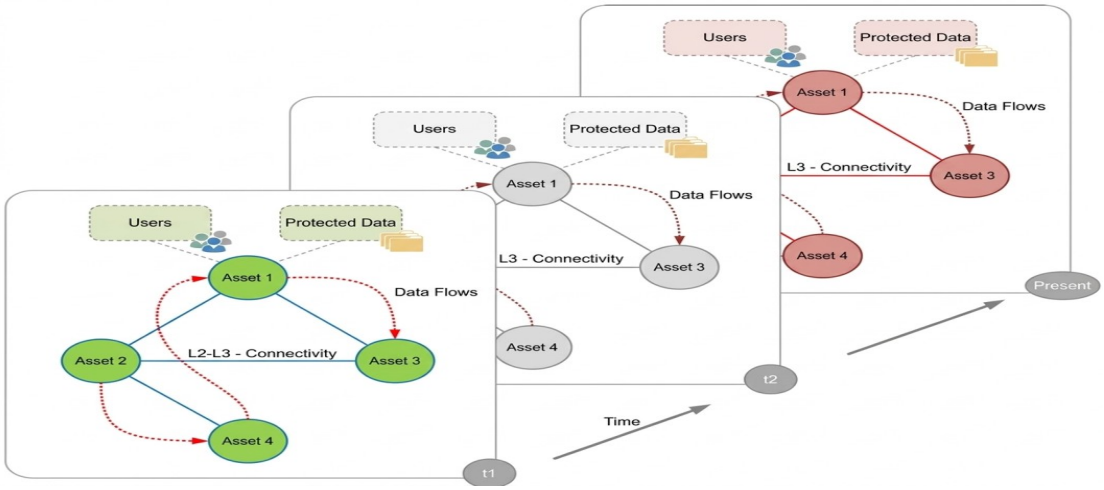


Figure 4: Evolution of the LAN feature model over time

Event aggregation groups related events and preserves source logs. Correlation rules define alerting combinations and frequencies; alternative SIEM approaches have been studied [7]. Thresholds and rules producing false positives require review under realistic loads.

Alert priority should reflect affected assets, accounts, and impact. NIST guidance links detection to incident response [12]. SIEM operations and alert filtering are discussed in [8–10]. Analysts can select representative attack techniques from MITRE ATT&CK [13], trace them in logs, and link distinguishing attributes to response actions. Figure 5 outlines this rule-development workflow.

5. Conclusions

This paper systematizes SIEM implementation principles linking source integration, current asset information, normalization, aggregation, and correlation-rule review. The result is a conceptual sequence for supporting incident analysis in an IT company.

The sequence has not been evaluated with operational security logs; reductions in false positives or response time therefore cannot be claimed. Future work should apply it to documented incident scenarios and compare alert quality and analyst workload with an existing SIEM configuration.

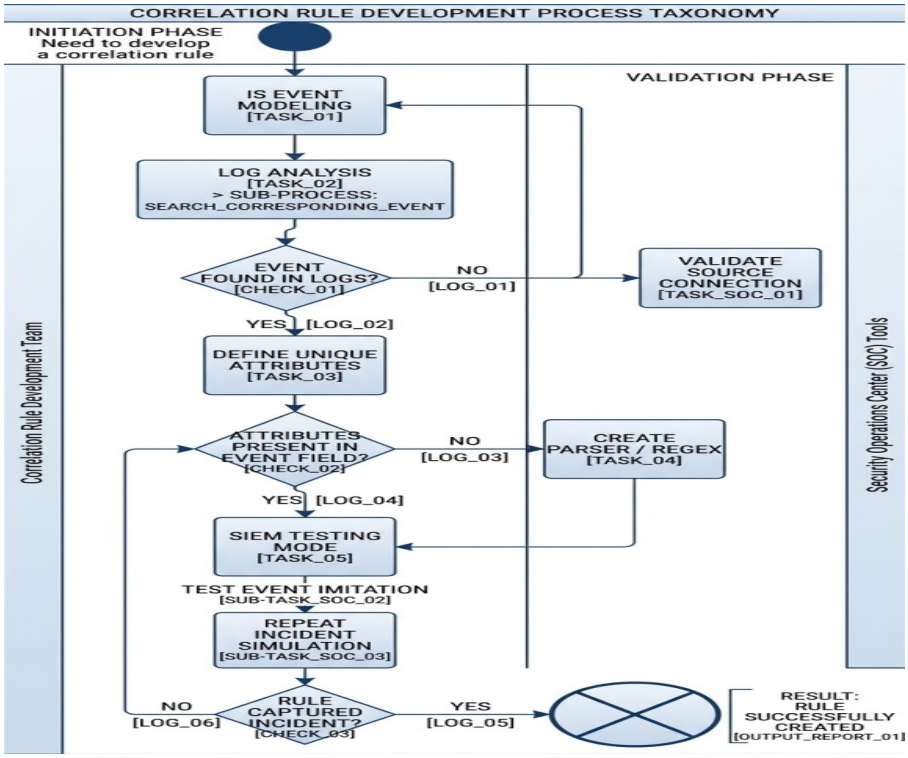


Figure 5: Workflow for developing and reviewing SIEM correlation rules.

Declaration on Generative AI

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) for text translation, paraphrasing and rewording, improving writing style, grammar and spelling checks, and formatting assistance. The authors critically reviewed and revised the suggested text, verified the scientific content and references, and take full responsibility for the final manuscript.

References

[1] A. Basta et al., “Security Information and Event Management (SIEM),” in Open-Source Security Operations Center (SOC), Wiley, 2025, doi: 10.1002/9781394201631.ch7.

[2] M. Cinque, D. Cotroneo, and A. Pecchia, “Challenges and Directions in Security Information and Event Management (SIEM),” Proc. IEEE ISSREW, 2018, pp. 95–99, doi: 10.1109/ISSREW.2018.00-24.

[3] G. González-Granadillo, S. González-Zarzosa, and R. Diaz, “Security Information and Event Management (SIEM): Analysis, Trends, and Usage in Critical Infrastructures,” Sensors, vol. 21, art. 4759, 2021, doi: 10.3390/s21144759.

[4] D. Díaz López et al., “Shielding IoT against Cyber-Attacks: An Event-Based Approach Using SIEM,” Wireless Communications and Mobile Computing, art. 3029638, 2018, doi: 10.1155/2018/3029638.

[5] K. Kent and M. Souppaya, Guide to Computer Security Log Management, NIST Special Publication 800-92, 2006, doi: 10.6028/NIST.SP.800-92.

[6] D. Jaeger et al., “Parallel and Distributed Normalization of Security Events for Instant Attack Analysis,” Proc. IEEE IPCCC, 2015, doi: 10.1109/PCCC.2015.7410270.

[7] G. González-Granadillo, M. El-Barbori, and H. Debar, “New Types of Alert Correlation for Security Information and Event Management Systems,” Proc. IEEE NTMS, 2016, doi: 10.1109/NTMS.2016.7792462.

[8] S. Bhatt, P. K. Manadhata, and L. Zomlot, “The Operational Role of Security Information and Event Management Systems,” IEEE Security & Privacy, vol. 12, no. 5, pp. 35–41, 2014, doi: 10.1109/MSP.2014.103.

[9] D. Cotroneo et al., “Empirical Analysis and Validation of Security Alerts Filtering Techniques,” IEEE Trans. Dependable Secure Comput., 16(5), 856–870, 2019, doi: 10.1109/ TDSC.2017.2714164.

[10] M. Cinque, R. Della Corte, and A. Pecchia, “Contextual Filtering and Prioritization of Computer Application Logs for Security Situational Awareness,” Future Generation Comput. Syst., 111, 668–680, 2020, doi: 10.1016/j.future.2019.09.005.

[11] Cybersecurity Insiders, 2022 SIEM Security Report, 2022. Available: https://static.fortra.com/core-security/pdfs/reports/cs-siem-report-2022.pdf/(accessed 2026-09-12).

[12] A. Nelson et al., Incident Response Recommendations and Considerations for Cybersecurity Risk Management, NIST SP 800-61 Rev. 3, 2025, doi: 10.6028/NIST.SP.800-61r3.

[13] MITRE, MITRE ATT&CK: Enterprise Tactics and Techniques. Available: https://attack.mitre.org/ resources/ (accessed 2026-09-22).

Empirical Validation of an Azure-Based Data Classification Pipeline for SOC 2 Type 2 Compliance: Benchmarking Eight Large Language Models

Oleh Deineka[1], Oleh Harasymchuk[1]

[1] *Lviv Polytechnic National University, 12 Bandera str., Lviv 79000, Ukraine*

Abstract

A previously proposed Azure-based framework for SOC 2 Type 2 information classification has been deployed as a pilot in a production-like Microsoft Fabric environment. This paper reports on what the deployment measured: whether metadata stays traceable and access control provably correct across the Bronze, Silver and Gold layers; which large language model should perform Silver-layer classification; and how far the pipeline's components cover SOC 2 Type 2 control objectives. Of eight models evaluated under an identical prompt and corpus, five reached 96.15% extraction quality while the cheapest reached only 39.6%, showing that per-token price is a poor proxy for suitability. Gemini 3 and Claude Sonnet 4.6 combined top-tier quality with the lowest latency, 6 and 7.5 seconds. Hallucination rates of 19 to 30 percent mean the output still requires rule-based verification.

Keywords

SOC 2 Type 2, Microsoft Fabric, Azure AI Foundry, Large Language Models, Medallion Architecture, Data Classification, Pilot Evaluation, Compliance Audit

1. Problem statement

SOC 2 Type 2 assesses not the design of an organization's controls but their operating effectiveness across an observation period of six to twelve months [1, 2]. Continuous monitoring, repeatable verification and traceable evidence therefore matter more than configuration quality at any single moment, and all five Trust Services Criteria depend on an auditable approach to classifying data and managing access to it.

The practical difficulty rarely lies in understanding the criteria but in operationalizing them across heterogeneous and rapidly growing volumes of structured, semi-structured and unstructured data held in collaboration platforms, lakes and relational stores at once [3]. Classification is normally maintained manually, so labels applied at creation are never revisited and the inventory diverges from reality within weeks. The commercial platforms that automate this work operate as closed systems: they expose dashboards and configuration but rarely document classification logic, decision thresholds or model behaviour in enough detail for an organization to verify or adapt them [4, 5].

2. Scientific novelty relative to known work

In earlier work we proposed an openly described framework for automated classification and secure storage of large corporate data volumes on Azure, built around Microsoft Fabric, Azure AI Foundry and a Medallion architecture, and compared it conceptually with commercial platforms [6]. Related studies examined the risks of LLM-based classification [7] and the detection of

Corresponding authors

deinekaoleg.86@gmail.com (O. Deineka); oleh.i.harasymchuk@lpnu.ua (O. Harasymchuk)

0009-0005-9156-3339 (O. Deineka); 0000-0002-8742-8872 (O. Harasymchuk)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

confidential information by language models [8], but both covered fewer models under differing protocols, and neither reported operation of a deployed pipeline.

The novelty of the present work is empirical rather than architectural. First, a re-executable verification protocol is applied to a working deployment, establishing metadata traceability across all three Medallion layers and provable correctness of access control, rather than asserting compliance by design. Second, eight models from three families are benchmarked under one identical prompt and corpus, with quality, latency, cost and context limit reported jointly, where earlier comparisons covered fewer models or optimized a single dimension. Third, pipeline components are mapped onto SOC 2 Type 2 control objectives using observed behaviour instead of vendor documentation, which makes the coverage claim falsifiable.

3. Proposed solution

SharePoint and OneDrive act as the data sources, covering both shared team repositories and the personal files that most often escape centralized governance. Fabric Data Factory authenticates against the Microsoft Graph API using a service-credentials flow with secrets held in Azure Key Vault, and moves data through three Medallion layers. Bronze stores raw metadata — identifiers, names, sizes, sharing flags, timestamps — as an inventory of information assets and a recovery point for audit. Silver applies cleansing and deduplication and delegates semantic classification of unstructured content to Azure AI Foundry, which performs entity extraction and tier assignment and returns PII, Finance and Customer records written as Delta tables. Gold aggregates the result for reporting through Power BI. The separation in Figure 1 between the data-processing plane and the control plane — identity, secrets, monitoring, alerting — is deliberate: security functions apply uniformly across the lifecycle rather than being attached to individual stages.

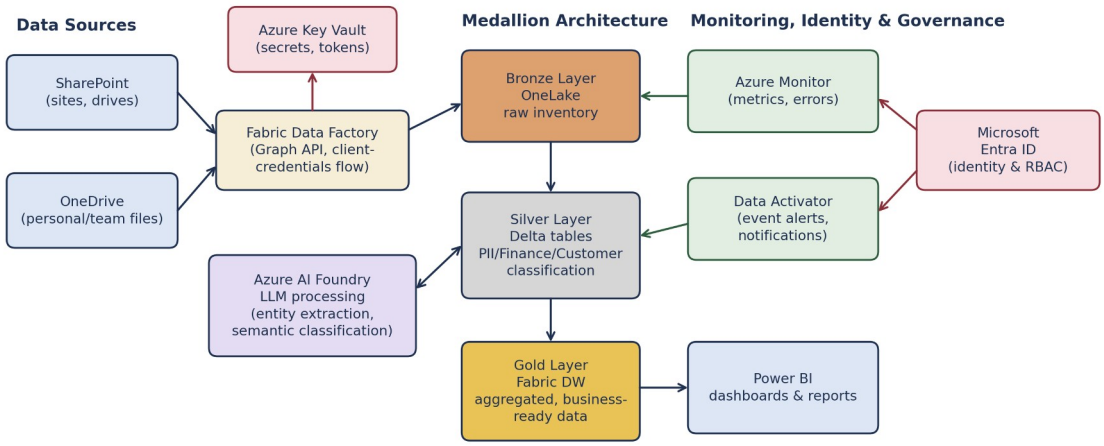


Figure 1: Pilot architecture. Data-processing plane.

4. Verification protocol

Integration testing validated the full path across both sources, Fabric Data Factory as orchestration, OneLake and the Fabric Data Warehouse as storage, and Azure AI Foundry as the classification engine. Automated scripts and manual spot-checks confirmed that document versions, authorship, timestamps and access rights were collected completely and stayed consistent from Bronze to Gold. Failure handling in the Graph API integration received particular attention: the pipeline refreshes

tokens on HTTP 401, retries with exponential backoff on 429 and 503, and records permanent failures without endless retries [2].

Access-control auditing verified that only authorized users could reach confidential data and reviewed encryption and key-storage practice. The pilot combined role-based access control with least privilege [9] and evaluated an attribute-based variant for cases where access depends on contextual attributes rather than static roles. Data Activator tracked encryption-policy violations and remediation timeliness, and multi-factor authentication was required for every privileged account touching the pipeline.

Confidentiality during processing was itself verified, since Silver-layer classification routes corporate content through third-party AI services. Corporate policy, Azure and Fabric access controls, and contractual data-handling commitments from the providers together proved sufficient, consistent with GDPR obligations where the data originates in the European Union [10].

5. LLM benchmark for sensitive-data detection

The central question for the Silver layer is which model best balances extraction quality, latency and cost when detecting PII, financial and customer data in unstructured documents. Eight models across three families were evaluated on an identical corpus with an identical prompt requesting structured JSON output across the three categories. Table 1 reports the results.

Table 1
Comparative evaluation of LLMs for automated sensitive-data detection

Model	Input, $/1K tok.	Output, $/1K tok.	Context limit	Quality, %	Time, s
GPT-3.5 Turbo	0.0031	0.0041	16K	80.77	24
GPT-4	0.0062	0.1236	32K	80.77	17
GPT-4o	0.0005	0.0150	128K	96.15	20
GPT-5.4	0.00025	0.0150	1M	96.15	25.5
Claude 3	0.0003	0.0150	200K	96.15	21
Claude Sonnet 4.6	0.0003	0.0015	1M	96.15	7.5
Gemini 1.5 Pro	0.000007	0.0003	128K	39.6	16
Gemini 3	0.00025	0.0150	1M	96.15	6

The results separate into three groups. The two earlier-generation OpenAI models plateaued at 80.77%, five newer models all reached 96.15%, and Gemini 1.5 Pro produced only 39.6% despite by far the lowest per-token cost, making it unsuitable for compliance-critical extraction regardless of price. Among the top-quality models, Gemini 3 and Claude Sonnet 4.6 combine that quality with the shortest response times, 6 and 7.5 seconds, at competitive cost.

Quality alone does not settle the choice. Context size determines whether long documents can be processed without chunking, and latency determines throughput. Because Bronze inventories hold hundreds of thousands of documents, a model that is dearer per call but needs fewer retries can be cheaper in aggregate, so model selection should be revisited periodically rather than fixed at deployment [7].

6. Control coverage

Because the pipeline was observed in operation, its effectiveness can be assessed component by component against explicit control objectives, as Table 2 summarizes.

Table 2
Mapping pipeline components to SOC 2 Type 2 control objectives

Component	Control objective	Pilot result	Impact
RBAC and least privilege [9]	Restrict unauthorized access	Role segregation and access-event auditing implemented	High
Azure AI Foundry classification	Detect PII, financial and customer data	Top-tier models reached 96.15% quality	High
Data Activator and event logging	Continuous monitoring and timely response	Event-driven notification of policy violations in place	High
Key Vault and encryption	Protect confidential data and secrets	Centralized secret management verified	High
Medallion layering	Traceability and integrity control	Auditability confirmed at each processing level	High
Backup and recovery	Availability and continuity	Recovery procedures in place; stable operation	Medium/High
Staff training	Reduce human-factor risk	Recurring regulated training planned	Medium

The components with the highest measured impact are those tied directly to access control, encryption, monitoring and audit — the areas SOC 2 Type 2 weighs most heavily. Together they form a control loop spanning inventory, classification, alerting and policy-compliance confirmation, which is what a Type 2 attestation examines.

7. Risks and limitations

Infrastructure dependency is the first constraint. A baseline architecture without redundancy typically targets around 99.5% availability, roughly 1.83 days of downtime per year, while 99.9% and 99.99% correspond to 8.77 hours and 52.6 minutes, each step costing materially more [11]. The second is hallucination: reported rates run from roughly 19% to 30% depending on the model, so classification output requires rule-based checks and selective expert review rather than unconditional trust [7]. The third is the human factor, which automation does not remove. A further limitation is scope: the pilot ran in a single organizational environment against a bounded set of SharePoint and OneDrive sources, so generalizing Table 1 to different document types, languages or regulatory contexts would require targeted revalidation.

8. Conclusions

Pilot testing confirmed complete metadata traceability across the three Medallion layers, correct enforcement of access-control and encryption policy, and reliable event-driven monitoring. The eight-model benchmark showed that Gemini 3 and Claude Sonnet 4.6 achieve the best combination of extraction quality, latency and cost, and that per-token price is a poor proxy for suitability: the cheapest model tested was also the least usable.

Component mapping showed high measured impact wherever a component touches access control, encryption or audit, and medium impact for training and recovery. Against commercial

platforms the pipeline trades a heavier in-house skill requirement for greater transparency and adaptability. Future work will track incident statistics across several audit cycles and carry out a total-cost-of-ownership analysis.

References

[1] AICPA. Illustrative Service Auditor's SOC 2 Type 2 Report. AICPA, 2022.
[2] AICPA. 2017 Trust Services Criteria for Security, Availability, Processing Integrity, Confidentiality, and Privacy (With Revised Points of Focus, 2022). AICPA, 2022.
[3] I. Gupta, A. K. Singh, C.-N. Lee, R. Buyya. Secure Data Storage and Sharing Techniques for Data Protection in Cloud Environments. IEEE Access, 2022.
[4] Varonis. SOC 2 Compliance, 2024. https://www.varonis.com/blog/soc-2-compliance
[5] Netwrix. Netwrix Auditor for SOC 2 Compliance, 2024. https://www.netwrix.com/compliance-audit-solution.html
[6] O. Deineka, O. Harasymchuk, A. Partyka. Comparative analysis of an Azure-based information classification methodology for SOC 2 Type 2. CEUR Workshop Proceedings, 2025.
[7] O. Deineka, O. Harasymchuk, A. Partyka, A. Obshta. Application of LLM for Assessing the Effectiveness and Potential Risks of the Information Classification System According to SOC 2 Type II. CEUR Workshop Proceedings, 2025, Vol. 3991, pp. 215-232.
[8] O. Deineka, O. Harasymchuk, A. Partyka, Y. Dreis, Y. Khokhlachova, Y. Pepa. Detection of confidential information using large language models. CEUR Workshop Proceedings, 2025.
[9] R. Sandhu et al. Role-based access control models. IEEE Computer, 1996, 29(2), pp. 38-47.
[10] Regulation (EU) 2016/679 (General Data Protection Regulation), 2016.
[11] K. Kaur, M. Singh. Fault Tolerance Techniques and Their Impact on Data Availability in Distributed Systems, 2021.

Cybersecurity: several major challenges

Mikolaj Karpinski[1], Szymon Czerkies[1], Magdalena Jankowicz[1], Michal Skowronek[1] and Mikolaj Stach[1]

[1] *University of the National Education Commission, Podchorazych 2, 30-084 Krakow, Poland*

Abstract

This paper discusses several major challenges in cybersecurity. In particular, it examines cybersecurity factors in the context of IT systems and IoT services. A close correlation was demonstrated between the level of device protection and global network availability. The tests carried out confirmed that the study objectives had been fully achieved.

Keywords

cyberattack, cybersecurity, IoT, network device, server infrastructure

1. Introduction

The last two decades have seen rapid development in information technology. The way many companies operate has changed. Digital transformation involves using digital technologies to modify virtually every aspect of an organisation's operations. In practice, this means moving away from traditional local systems. Companies are increasingly opting for distributed solutions operating in a cloud-based model [1].

The cloud computing model enables organisations to utilise IT resources in a flexible and scalable manner, via a network providing access to a shared pool of resources (e.g. applications and services, servers, networks), which can be easily made available by the management team or by the service provider [2].

The cloud model comprises three service models: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [3]. Each type is characterised by a different level of responsibility between the customer and the provider.

Network security encompasses not only the cryptographic protection of information in transit, but also the network devices themselves. The starting point for security modelling is the CIA triad, which consists of confidentiality, integrity and availability [4].

The NIST SP 800-30 standard describes the terminology used in assessing the effectiveness of information security measures [5].

The evolution of the methods used by cybercriminals means that network administrators must constantly adapt their defence strategies. As indicated in the European Union Agency for Cybersecurity (ENISA) Threat Landscape Report, cyberattacks are becoming increasingly sophisticated [6].

Given the ineffectiveness of centralised trust models, the most recent literature on the subject from 2024–2026 places strong emphasis on the search for innovative, decentralised access control mechanisms. One of the most promising areas of research is the integration of IoT architecture with blockchain technology [7]. The use of so-called smart contracts, transferred from the cloud directly to the edge computing layer, enables the creation of immutable, fully auditable

Corresponding author

mikolaj.karpinski@uken.krakow.pl (M. Karpinski)

0000-0002-8846-332X (M. Karpinski)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

authorisation logs [7, 8]. In this paradigm, a device's identity and its authorisations are cryptographically verified by a distributed network of edge nodes, which eliminates the need to rely on 'heavy' PKI servers and effectively reduces the risk of a single point of failure (SPOF) in the area of authentication [7].

The aims of this work are: to demonstrate a close correlation between the level of device protection and global network availability; to analyse cybersecurity factors in the context of IT systems, and to assess their significance for the protection of the cloud environment of a company operating in the e-commerce sector; and to conduct a multidimensional analysis of the security architecture of IoT systems.

2. Materials, methods and results

The conclusions drawn from the theoretical analysis indicate that Control Plane Policing (CoPP) is one of the key and most effective mechanisms for combating DoS attacks on network devices. It enables precise, selective protection of the processor, which is a key prerequisite for maintaining network continuity and high performance. It has been demonstrated that isolation mechanisms, such as CoPP, are critical for maintaining CPU availability, whilst the implementation of strong cryptography (SNMPv3, IPSec) protects the management and data planes, but introduces a measurable performance overhead.

The phishing simulation was carried out under controlled organisational conditions as a one-off campaign. The study involved a group of 83 employees of the organisation. This group was selected purposefully and comprised users who regularly use email and organisational systems accessible via the identity management environment. The message was successfully delivered to all recipients. It is worth noting that the study was conducted in accordance with organisational and ethical safety principles. This simulation was intended solely for training and analytical purposes.

A multidimensional empirical analysis – covering both privacy-oriented consumer ecosystems (Smart Home) and deterministic industrial networks (IIoT) – unequivocally demonstrates that traditional, perimeter-based protection models are wholly inadequate when confronted with the realities of the IoT. The deployment of firewalls at the LAN/WAN interface or network segmentation using virtual private networks (VPNs) merely masks fundamental shortcomings in firmware security. To effectively neutralise identified attack vectors, security engineering must evolve towards a holistic "Security-by-Design" approach, in which protection is implemented from the lowest silicon layer right up to the cloud API.

A thorough assessment of IoT ecosystems, based on global security repositories – and in particular the National Vulnerability Database (NVD) – provides irrefutable evidence that the current model for the manufacture and maintenance of embedded devices may appear to be in a state of systemic crisis. The aggregation and quantification of recorded vulnerabilities (CVEs) from 2024 to 2026 allows us to draw significant conclusions regarding the nature of the threats, which have evolved from isolated software vulnerabilities towards structured attack vectors targeting entire network architectures.

The first and most striking finding to emerge from the analysis of the CVE database is a marked disparity between the level of security maturity in traditional IT systems and the standards prevailing in the IoT sector. In the case of consumer devices (Smart Home), the statistics reveal that manufacturers are completely disregarding the principles of the Secure Software Development Lifecycle (SSDLC). The dominant category of reports in the NVD database for this sector remains firmware memory errors, with a particular emphasis on buffer overflow vulnerabilities. Cross-referencing this data with the prevalence of hardcoded credentials leads to the conclusion that the majority of low-cost sensors and IP cameras are, by design, incapable of fending off even primitive

automated attacks. The result of this phenomenon is the ongoing resurgence of botnets derived from the Mirai family, for which consumer devices represent an inexhaustible source of computing resources.

A completely different, yet equally worrying picture emerges from an analysis of vulnerabilities in the IIoT sector and in medical systems (IoMT). A review of the CVE database in the context of internal infrastructure shows that historical technological debt and the ongoing convergence of IT and operational technology (OT) environments constitute internal risk factors.

3. Conclusions

In the unprotected phase, DoS attacks caused immediate CPU performance collapse (>99%), leading to the instability of dynamic routing protocols (OSPFv2 flapping). The implementation of the Control Plane Policing (CoPP) mechanism proved to be an effective line of defense against low- and medium-intensity attacks, reducing CPU load to a safe level and maintaining network stability. However, a critical finding was the Self-DoS phenomenon, caused by the incorrect aggregation of management (SSH) and diagnostic (ICMP) traffic within a single CoPP class. The subsequent optimization (CoPP v2) with full traffic separation eliminated this vulnerability, confirming the necessity of granular configuration. It has been proven that CoPP is a vital strategy against communication paralysis, and network engineers should adopt best practices, including class separation and secure management protocols (SSHv2, SNMPv3, TACACS+). The physical limitations of software-based protection against massive volumetric attacks (receive livelock phenomenon) have also been identified, underscoring the necessity of dedicated hardware (ASIC) for complete resilience.

To conclude the discussions undertaken in the present study of IoT solutions, which aims to emphasise unequivocally that this technology has definitively transcended the realm of digital convenience, becoming the nervous system of the modern economy and internal infrastructure. Building resilience in such a complex environment is not a task that can be solved by a single protocol. The guarantee of tomorrow's security lies in the holistic synergy between attack-resistant hardware, cryptography adapted to physical constraints, rigorous isolation of network traffic, and autonomous, edge-based threat analytics. Only an ecosystem designed in this way will allow the full potential of M2M networks to be realised, whilst protecting users' privacy and the physical integrity of the world around them.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] E. Ghazaryan, Cloud Computing as a Catalyst for Digital Transformation in Enterprises, The American Journal of Engineering and Technology, 7 (2025) 170–177. doi: 10.37547/tajet/Volume07Issue06-19.

[2] P. Mell, T. Grance, The NIST Definition of Cloud Computing, Special Publication 800-145, National Institute of Standards and Technology, Gaithersburg, 2011.

[3] Narodowe Standardy Cyberbezpieczenstwa. Standardy Cyberbezpieczenstwa Chmur Obliczeniowych SCCO, wersja 2.00. Ministerstwo Cyfryzacji, Warszawa, 2025.

[4] ISO/IEC, ISO/IEC 27000: 2018. Information technology — Security techniques — Information security management systems — Overview and vocabulary, ISO/IEC, 2018.

[5] Joint Task Force, Guide for Conducting Risk Assessments. NIST Special Publication 800-30 Revision 1, National Institute of Standards and Technology, 2012, Appendix B: Glossary. URL: https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-30r1.pdf.
[6] European Union Agency for Cybersecurity, ENISA Threat Landscape 2024, European Union Agency for Cybersecurity, 2024. URL: https://www.enisa.europa.eu/sites/default/files/2024-
[7] S. Sadhwani, U.K. Modi, R. Muthalagu, P.M. Pawar, SmartSentry: Cyber Threat Intelligence in Industrial IoT, IEEE Access 12 (2024), 34720-34740. doi: 10.1109/ACCESS.2024.3371996.
[8] M. Almutairi, F.T. Sheldon, IoT-Cloud Integration Security: A Survey of Challenges, Solutions, and Directions, Electronics 14 (2025). doi: 10.3390/electronics14071394.
[9] iperf3 — official website, 2026. URL: https://iperf.fr/.

Set-Theoretic Approach to Automating Cyber-Resilience Assessment of Critical Infrastructure Facilities

Yuliia **Khokhlachova**[1,2], Kateryna **Khokhlachova**[3]

[1] *State University of Information and Communication Technologies, 7 Solomianska St., Kyiv, 03110, Ukraine*
[2] *State University of Trade and Economics, 19 Kyoto St., Kyiv, 02156, Ukraine*
[3] *Taras Shevchenko National University of Kyiv, 24 Bohdana Havrylyshyna St., Kyiv, 04116, Ukraine*

Abstract

The accelerating sophistication and impact of adversarial cyber campaigns against Critical Infrastructure Facilities (CIF) necessitate a fundamental shift from qualitative perimeter defense to quantifiable, automated cyber-resilience assessment. Existing conceptual guidelines, such as the MITRE Cyber Resiliency Engineering Framework (CREF) and NIST standards, offer comprehensive taxonomies but lack formal mathematical mechanisms to map baseline protective activities to verifiable metrics and computed integral resilience indices. This paper develops a formalized Set-Theoretic Metric Model (STMM) designed to automate multi-tier CIF resilience evaluation. The model establishes set-theoretic functional mappings across baseline engineering measures and techniques, introduces a deterministic 5-index hierarchical classification, defines metric cardinality (ranging from 1 to 11), and harmonizes ex-ante architectural readiness with ex-post operational telemetry. The algorithmic pipeline is validated using a spear-phishing attack scenario against a municipal core banking facility, proving STMM's utility for real-time SOC integration and digital twin synchronization.

Keywords

Cyber-resilience, Critical Infrastructure, Set-Theoretic Metric Model, MITRE CREF, Metric Cardinality, Automated Assessment, Ex-Ante/Ex-Post Telemetry.

1. Introduction

Modern Critical Infrastructure Facilities (CIF) – encompassing energy grids, nuclear stations, water supply networks, telecommunication nodes, and financial clearing complexes – operate in an environment of accelerating digital threat vectors. The deep convergence of Industrial Internet of Things (IIoT) sensors, SCADA platforms, and distributed cloud architectures expands the operational attack surface while creating tight inter-system couplings. Consequently, adversarial campaigns – ranging from Advanced Persistent Threats (APTs) and zero-day exploits to supply-chain poisoning and destructive ransomware – can trigger catastrophic cascading failures across interdependent civil and national infrastructure ecosystems.

Traditional cybersecurity doctrines anchored in perimeter-centric prevention, barrier isolation, and static intrusion detection have proven structurally inadequate where system compromise must realistically be presumed inevitable. In response, cyber-resilience has emerged as the definitive engineering paradigm: the adaptive property of a mission system to anticipate, withstand, recover from, and evolve under hostile cyber conditions, active stresses, and compromises while sustaining critical delivery functions.

Notwithstanding its recognized strategic necessity, operationalizing and automating cyber-resilience assessment in practical CIF environments remains hindered by methodological fragmentation. Leading architectural frameworks, notably MITRE CREF and NIST SP 800-160, delineate rich multi-tier hierarchies of goals and techniques; however, they remain qualitative and

Corresponding author
khohlachova@gmail.com (K. Khokhlachova)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

descriptive. Security practitioners lack formal mathematical apparatuses to connect baseline defensive actions with verifiable metric sets, quantify metric cardinality, differentiate between proactive readiness and runtime telemetry, or compute reproducible integral indices. Subjective checklist audits introduce human bias and prevent real-time SOC automation. To overcome these deficiencies, this study develops a Set-Theoretic Metric Model (STMM) establishing a mathematically unified, algorithmically verifiable measurement pipeline.

2. State of the Art and Comparative Analysis

A rigorous review of academic literature, engineering standards, and international regulations highlights persistent gaps across four foundational criteria: formalization of hierarchical structures (C1), establishment of rigorous mathematical dependencies (C2), metric indexing and cardinality quantification (C3), and integrated threat-to-metric mappings (C4).

Framework / Research Source	C1: Hierarchy Formalization	C2: Mathematical Dependencies	C3: Indexing & Cardinality	C4: Threat-to-Metric Mapping
MITRE CREF Taxonomy [1–3]	Descriptive / High	Very Low (Heuristic)	Absent / Informal	Low (Qualitative Tables)
NIST SP 800-160 & CSF 2.0 [4, 9]	Medium-Low	Low (Checklists)	Absent	Absent
EU DORA Directives [10]	Organizational	Absent	Absent	Absent (Audit-Based)
Mission Impact & SCADA Models [6, 7]	Domain-Specific	Medium (Simulation)	Absent	Low (Scenario Bounds)
General Resilience Metric Studies [5, 8]	Low	Medium-High	Fragmented	Low
Proposed STMM Framework	**Complete Set-Theoretic**	**Functional Mappings**	**Rigorous 5-Index System**	**Integrated (Ex-Ante/Ex-Post)**

The comparative synthesis indicates that prior research focuses either on qualitative compliance or domain-specific simulations without delivering an automated mathematical architecture that links discrete protective actions to verifiable telemetry metrics.

3. Mathematical Formalization of the Set-Theoretic Metric Model (STMM)

To enable automated, reproducible evaluation, the STMM structures the quantitative resilience measurement domain across four hierarchical tiers:

Tier 1: Strategic Goals (GS). The global resilience objective space is defined as $GS = \{g_1, g_2, ..., g_n\}$, where $n = |GS| = 4$, representing Anticipate (ANT), Withstand (WITH), Recover (REC), and Adapt (AD).

Tier 2: Tactical Objectives (OS). Operationalizing strategic goals requires tactical objectives decomposed into specific sub-objectives: OS = {OS_1, OS_2, ..., OS_m}, with m = |OS| = 8: Prevent/Avoid (PA), Prepare (PR), Continue (CN), Constrain (CS), Reconstitute (RE), Understand (UN), Transform (TR), and Re-architect (RA). Each objective decomposes into s_i sub-objectives: OS_i = {OS_{i,1}, OS_{i,2}, ..., OS_{i,s_i}} for i ∈ {1, ..., m}.

Tier 3: Parameterized Baseline Measures (BM). Sub-objectives OS_{ij} are executed via functional baseline measures parameterized by technical implementation methods A_{ijkl}:

$$OS_{ij} = \{ OS_{ij,1}(A_{ij11}, ..., A_{ij1,a_{ij1}}), ..., OS_{ij,k_{ij}}(A_{ijk1}, ..., A_{ijk,a_{ijk}}) \},$$

where k_{ij} is the number of baseline activities under sub-objective j, and a_{ijk} represents the parameter arity of implementation methods: Analytic Monitoring (AM), Contextual Awareness (CA), Substantiated Integrity (SI), Privilege Restriction (PR), Redundancy (R), and Coordinated Protection (CP).

Tier 4: Metric Sets and Cardinality (MMM). For each baseline measure OS_{ijk}(A_{ijk1}, ..., A_{ijk,a_{ijk}}), the model defines a corresponding verifiable metric set M_{ijk}:

$$M_{ijk} = \{ M_{ijk,1}(A_{ijk1,1}, ..., A_{ijk1,a_{ijk1}}), ..., M_{ijk,q_{ijk}}(A_{ijkq,1}, ..., A_{ijkq,a_{ijkq}}) \},$$

where q_{ijk} = |M_{ijk}| denotes metric set cardinality. The complete model is formalized as the grand Cartesian union:

$$STMM = \cup_{i=1}^{m} \cup_{j=1}^{s_i} \cup_{k=1}^{k_{ij}} (OS_{ijk}(A_{ijk1}, ..., A_{ijk,a_{ijk}}), \cup_{q=1}^{q_{ijk}} M_{ijk,q}(A_{ijkq,1}, ..., A_{ijkq,a_{ijkq}})).$$

Across 114 baseline actions in an enterprise CREF deployment, metric cardinality ranges deterministically from 1 to 11 based on technical complexity: q_{ijk} = 1 for singular focused tasks (e.g., PA.S1.A4, RE.S4.A4); q_{ijk} = 2–3 for structural verifications (e.g., PA.S1.A2, TR.S1.A1); and q_{ijk} ≥ 4 for complex telemetry and integrity validation (e.g., CN.S1.A1 with q=10, CN.S2.A1 with q=10, RE.S1.A2 with q=6, and UN.S1.A1 with q=6).

Crucially, STMM bifurcates indicators into two complementary categories: (1) Structural Metrics (M^{struct}, Ex-Ante) evaluating baseline posture, configuration hygiene, and resource isolation prior to intrusions; and (2) Operational Metrics (M^{oper}, Ex-Post) capturing dynamic telemetry, latency, containment, and transaction verification during and following an incident.

Index (i.j.k.q)	Identifier	Techniques	Classification	Operational Interpretation & Data Source
1.1.1.1	PA.S1.A1.M1	PR, S	Structural (Ex-Ante)	Ratio of cyber assets governed by RBAC/ABAC. Source: IAM directory logs.
1.1.3.3	PA.S1.A3.M3	S, R	Structural (Ex-Ante)	Cross-enclave discoverability ratio across segmented subnets. Source: Telemetry.
3.1.1.1	CN.S1.A1.AM-DA-1	AM, CA, SI	Operational (Ex-Post)	Mean elapsed time to quantify mission impact (h). Source: SIEM alert logs.

Index (i.j.k.q)	Identifier	Techniques	Classification	Operational Interpretation & Data Source
3.1.1.4	CN.S1.A1.M1	AM, CA, SI	Operational (Ex-Post)	Percentage of databases with verified cryptographic integrity post-CCoA.
5.1.2.3	RE.S1.A2.RE-S1-A1-1	AM, SI	Operational (Ex-Post)	Latency to detect corrupt records in transactional storage (min). Source: DB logs.
7.2.2.2	TR.S2.A2.M2	R, CP	Structural (Ex-Ante)	Ratio of remediated excessive privilege assignments. Source: PAM audit.

4. Operationalization and Scenario-Based Validation

To demonstrate the model's capacity for automated algorithmic assessment, STMM was validated against a targeted spear-phishing attack scenario on a municipal core banking entity. Unlike volumetric flood attacks, phishing targets access tokens, human trust, and transaction pipelines without immediate physical node disruptions.

4.1. Tactical Subgraph Activation

Under the spear-phishing scenario, the model instantiates an active measurement subgraph across all strategic goals:

- Goal **ANT** (g_1): Deploys objective UN. Sub-objective UN.S1 instantiates 6 threat intelligence metrics: active feeds (M1 = 4), refresh frequency (M2 = 15 min), relevance assessment delay (M3 = 25 min), TTP publication delay (M4 = 45 min), intelligence sharing rate (M5 = 1/week), and threat groups tracked (M6 = 4). Sub-objective UN.S2 executes 3 metrics: Mission Impact Analysis (MIA) update age (M1 = 490 days), MIA verification status (M2 = verified), and critical cyber assets ratio (M3 = 80%).
- Goal **WITH** (g_2): Deploys objective CN. Sub-objective CN.S1 executes 10 metrics: damage assessment latency (AM-DA-1 = 3.5 h), cryptographic check latency (SI-IC-1 = 18 min), verifiable critical data ratio (SI-IC-2 = 85%), verified assets post-CCoA (M1 = 70%), log audit latency (SI-IC-3 = 25 min), log integrity ratio (SI-IC-4 = 90%), service check latency (SI-BV-1 = 40 min), verified business applications (M2 = 75%), security daemon check latency (SI-BV-2 = 55 min), verified security applications (M3 = 75%), and audit frequency (SI-IC-5 = 1 check per 6 h). CN.S2 executes 10 parallel telemetry metrics.
- Goal **REC** (g_3): Deploys objective RE. Sub-objective RE.S1 initiates 6 metrics: critical data verification (SI-IC-2 = 90%), supporting data ratio (SI-IC-5 = 85%), anomalous record detection latency (RE-S1-A1-1 = 18 min), active integrity checkpoints (CS-S1-A1-2 = 4), schema validation conformance (SI-IC-6 = 100%), and executive notification time (RE-S1-A1-2 = 30 min).
- Goal **AD** (g_4): Deploys objective TR. Sub-objective TR.S1 tracks Single Point of Failure (SPOF) elimination (analyzed flows = 80%; SPOF-free flows = 60%). TR.S2 tracks privilege remediation (audited accounts = 80%; revoked over-privileged roles = 75%).

4.2. Algorithmic Data Ingestion and Aggregation

Rather than relying on manual questionnaires, STMM enables direct ingestion via SIEM/SOAR APIs, Identity Providers (Okta, Entra ID), and EDR platforms. The computational pipeline executes three steps:

1. Telemetry Harvesting: Continuous streaming of log events, cryptographic hashes, and timestamps.
2. Normalization: Metric values with disparate units (latencies in hours vs. dimensionless ratios) are mapped into [0, 1] using SLA-bounded piecewise utility functions: μ(M_delay) = max(0, min(1, (T_max - t)/(T_max - T_target))), μ(M_ratio) = x / 100%.
3. Hierarchical Synthesis: Normalized metrics are synthesized using multi-criteria priority weight vectors (w) derived from the Analytic Hierarchy Process (AHP) and entropy weighting:

$$R_{CIF} = \Sigma_{i=1}^{m} w_i \Sigma_{j=1}^{s_i} w_{ij} \Sigma_{k=1}^{k_{ij}} w_{ijk} \Sigma_{q=1}^{q_{ijk}} w_{ijkq} \cdot \mu(M_{ijk,q}),$$

where weight vectors satisfy simplex constraints Σ w = 1 at all hierarchical levels, yielding an objective integral index $R_{CIF} \in [0, 1]$.

5. Architectural Discussion and Implementation Impact

Implementing STMM provides four critical advantages over conventional assessment mechanisms:

1. Elimination of Subjectivity: By coupling each metric to specific technical methods (AM, CA, SI, R, CP), subjective scoring is replaced with immutable digital artifacts, cryptographic proofs, and SIEM timestamps.
2. Dynamic Cardinality Management: The framework scales smoothly across maturity tiers; advanced SOCs can evaluate all operational telemetry metrics, while emerging utilities can begin with baseline structural indicators without modifying model syntax.
3. Cyber Resilience Digital Twins (CRDT): STMM provides a formalized schema for synchronizing real-time telemetry with virtual infrastructure twins, enabling stress-testing of proactive Courses of Action (CCoA).
4. Regulatory Compliance Automation: Direct structural traceability supports compliance pipelines for international mandates, including the EU DORA regulation and NIS2 Directive, replacing annual audits with continuous monitoring.

6. Conclusions

This paper presents the Set-Theoretic Metric Model (STMM), resolving the lack of mathematical formalization in critical infrastructure cyber-resilience assessment. The primary contributions comprise: (1) formal set-theoretic mapping linking strategic goals, tactical objectives, baseline measures, and quantitative metrics; (2) deterministic quantification of metric cardinality (ranging from 1 to 11) within an unambiguous 5-index hierarchy (i.j.k.q); (3) harmonization of ex-ante structural readiness and ex-post incident telemetry; and (4) operational validation under a banking spear-phishing scenario demonstrating automated telemetry ingestion and index synthesis. Future work will focus on autonomous machine learning for dynamic weight recalibration during zero-day conditions and field deployment across municipal distributed energy microgrids.

References

[1] Bodeau D., Graubart R., McQuaid R., Woodill J. Cyber Resiliency Metrics, Measures of Effectiveness, and Scoring. MITRE Technical Report MTR180314, The MITRE Corporation, Bedford, MA, 2018.

[2] Bodeau D., Graubart R. Cyber Resiliency Engineering Framework. MITRE Corporation Technical Report MTR110237, Bedford, MA, 2011.

[3] Bodeau D., Brtis J., Graubart R., Salwen J. Cyber Resiliency Engineering Aid. MITRE Technical Report MTR150264, Bedford, MA, 2015.

[4] Ross R., Pillitteri V., Graubart R., Bodeau D., McQuaid R. Developing Cyber-Resilient Systems: A Systems Security Engineering Approach. NIST SP 800-160, Vol. 2, Rev. 1, 2021.

[5] Linkov I., Eisenberg D. A., Plourde K., et al. Resilience metrics for cyber systems. Environment Systems and Decisions, Vol. 33, No. 4, 2013, pp. 471–476.

[6] Musman S., Temin A. A cyber mission impact assessment tool. In: Proc. IEEE HST, Waltham, MA, 2015, pp. 1–7.

[7] Falco G., Caldera C., Shrobe H. IIoT cybersecurity risk modeling for SCADA systems. IEEE Internet of Things Journal, Vol. 5, No. 6, 2018, pp. 4486–4495.

[8] Haque M. A., Shetty S., Krishnappa B. Cybersecurity resilience model for industrial control systems. IET Cyber-Physical Systems, Vol. 5, No. 3, 2020, pp. 271–282.

[9] National Institute of Standards and Technology (NIST). The NIST Cybersecurity Framework (CSF) 2.0. NIST CSWP 29, Gaithersburg, MD, 2024.

[10] European Parliament and Council. Regulation (EU) 2022/2554 on digital operational resilience for the financial sector (DORA). Official Journal of the European Union, L 333, 2022.

[11] AlHidaifi A., et al. Simulation-Based Quantification of Cyber Resilience for Critical Systems. Decision Support Systems, Vol. 178, 2024.

[12] Cho H., et al. A Quantitative Framework for Cyber Resilience Assessment with Normalization Techniques. Electronics, Vol. 14, No. 12, 2025, Art. no. 2465.

Mathematical Method and Criteria for Assessing the Physical Security of Server Hardware in Cloud Services

Yevhenii Pedchenko[1], Yevheniia Ivanchenko[1], Ihor Ivanchenko[1], Maryna Pedchenko[1] and Yurii Korovaichenko[1]

[1] *State University of Information and Communication Technologies, Solomianska str. 7, 03110 Kyiv, Ukraine*

Abstract

This paper presents the structure and mathematical method of the fourth component of a previously developed generalized mathematical model for assessing the cyber defense state of cloud services of information infrastructure objects – the server hardware module (parameter SR, "Server hardware module"), which quantitatively determines the level of physical protection of the physical servers used to deliver a cloud service. Based on a set-theoretic approach, the parameter SR is formalized as a tuple of twelve basic criteria, each of which is further decomposed into nested subsets of answer options, covering the level of physical access to servers, video surveillance, the method of access control to the data center, uninterruptible power supply, climate control, fire-suppression systems, motion/alarm/identity-verification systems, the equipment refresh cycle, next-generation firewalls, intrusion detection and prevention systems, the possibility of unauthorized physical connection to servers, and the data-recovery behavior in the event of a physical server failure. The generalized formulas describing the SR module and its criteria and sub-criteria are presented, together with a summary table of all twelve criteria and their answer options. The presented information allows auditors to obtain a quantitative, well-grounded assessment of the physical security of a cloud service provider's server infrastructure prior to the decision on its use in the company's production environment.

Keywords

cybersecurity, information security, cloud service, physical security, model, assessment method, audit

1. Introduction

The mass adoption of cloud computing is accompanied by a steady increase in the number of cyber incidents affecting the operability of cloud infrastructure. Modern industry recommendations are focused on the network level (DDoS mitigation, segmentation, the use of secure access protocols) and on the data storage level (encryption, key management, compliance with regulatory requirements for data protection) [1]. However, considerably less attention is paid to the physical level of cloud infrastructure – the data centers that house the physical servers on which virtualized or containerized cloud services ultimately run [2].

At the same time, breaches of physical security remain a root cause of service disruptions and data leaks: unauthorized access to premises, insufficient power-supply redundancy, failures of microclimate control systems, and the absence of a controlled equipment decommissioning process can compromise confidentiality, integrity and availability no less effectively than an ordinary cyberattack, and in some cases completely bypass network- and application-level protection measures [3]. The presence of such risks calls for the inclusion, within the generalized model for assessing the cyber defense of cloud services, of a separate and clearly defined module for assessing the physical protection of server hardware [4].

Corresponding authors

e.pedchenko@duikt.edu.ua (Y. Pedchenko); e.ivanchenko@duikt.edu.ua (Y. Ivanchenko); i.ivanchenko@duikt.edu.ua (I. Ivanchenko); st9763427@duikt.edu.ua (M. Pedchenko); y.korovaichenko@duikt.edu.ua (Y. Korovaichenko)

0000-0001-8436-5792 (Y. Pedchenko); 0000-0003-3017-5752 (Y. Ivanchenko); 0000-0003-3415-9039 (I. Ivanchenko); 0009-0007-8743-1821 (M. Pedchenko); 0009-0002-4065-5776 (Y. Korovaichenko)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Analysis of Recent Research and Publications

Industry frameworks for the protection of the physical infrastructure of data centers include the Uptime Institute Tier Standard [3] and the ANSI/TIA-942 Telecommunications Infrastructure Standard for Data Centers [4], which define a tiered classification of redundancy, fault tolerance and physical protection of power, cooling and access-control subsystems and are widely used for the certification of facilities and data centers. The ISO/IEC 27001:2022 standard, in Annex A, groups physical security requirements into the "physical controls" category (controls A.7.1-A.7.14), covering physical security perimeters, entry control, protection against environmental threats, equipment siting, and secure equipment disposal [5]. NIST SP 800-53 Rev. 5 defines an analogous set of "Physical and Environmental Protection" (PE) controls, covering physical access authorization, monitoring, power and cabling systems, and emergency power and lighting [6]. The PCI-DSS standard likewise requires physical access control and video monitoring of premises where payment card data is stored [7].

Although these frameworks provide ready-made checklists and tiered certification classifications, none of them offers a formalized quantitative model that would allow an auditor to determine the physical security state of a specific cloud service provider on a per-criterion basis, comparable across different providers and combinable with other cyber-defense assessment parameters. For the assessment of cloud providers, the use of a generalized tuple-based mathematical model for assessing the cyber defense state of cloud services, CSP = {GP, N, S, SR, V, OS, CT, R, A, D, RE}, has been proposed [8]. Within the scope of this research, the fourth component of the tuple – the server hardware module SR – is described [8].

The aim of this research is to formalize the structure and the set-theoretic mathematical method of the server hardware module (parameter SR), based on the previously developed generalized model for assessing the cyber defense of cloud services, by defining its basic criteria and sub-criteria required for the quantitative assessment of the level of physical protection of a cloud service provider's server infrastructure.

3. Results

The server hardware module (parameter SR, referred to in [8] as the "Server hardware module") is the fourth component of the generalized tuple CSP = {GP, N, S, SR, V, OS, CT, R, A, D, RE}. Its purpose is to quantitatively determine the level of physical protection of the server hardware used for assessing the security level of a cloud service, independently of the other assessment parameters [9], [10]. In accordance with the set-theoretic approach adopted for the whole model, the parameter SR is defined by Formula 1:

$$\mathbf{SR} = \left\{ \bigcup_{j=1}^{m_4} \mathbf{SR}_j \right\} = \left\{ \mathbf{SR}_1, \mathbf{SR}_2 \ldots, \mathbf{SR}_{m_4} \right\}, \quad (1)$$

where: $\mathbf{SR}_j \subseteq \mathbf{SR}$ $(j = \overline{1, m_4})$ – describes the j-th criterion for assessing the physical protection of the server hardware of a cloud service provider, and m_4 – denotes their number.

At $m_4 = 12$, Formula 1 takes the form:

$$\mathbf{SR} = \left\{ \bigcup_{j=1}^{12} \mathbf{SR_j} \right\} = \tag{2}$$

$$\left\{ \mathbf{SR_1}, \mathbf{SR_2}, \mathbf{SR_3}, \mathbf{SR_4}, \mathbf{SR_5}, \mathbf{SR_6}, \mathbf{SR_7}, \mathbf{SR_8}, \mathbf{SR_9}, \mathbf{SR_{10}}, \mathbf{SR_{11}}, \mathbf{SR_{12}} \right\},$$

where, respectively: $\mathbf{SR_1}$ = “Select the level of physical access to the servers (Hardware) of the cloud service:”; $\mathbf{SR_2}$ = “The physical equipment of the cloud service is under continuous video surveillance:”; $\mathbf{SR_3}$ = “Access to the data center premises is granted by:”; $\mathbf{SR_4}$ = “The data center provides an uninterruptible power supply even in the event of a blackout:”; $\mathbf{SR_5}$ = “The data center provides a reliable climate control system for the premises housing the cloud service provider’s servers:”; $\mathbf{SR_6}$ = “All premises in the data center are equipped with fire-suppression systems:”; $\mathbf{SR_7}$ = “The data center has motion detectors, alarm systems and identity-verification systems in every room where equipment is located:”; $\mathbf{SR_8}$ = “The equipment in the data center is refreshed every 4 years to ensure fault tolerance and improve equipment performance:”; $\mathbf{SR_9}$ = “Access to the cloud service’s servers is protected by next-generation firewalls (NGFW):”; $\mathbf{SR_{10}}$ = “Access to the cloud service’s servers is monitored by intrusion detection and prevention systems (IPS & IDS):”; $\mathbf{SR_{11}}$ = “Physical connection to the cloud service provider’s servers:”; $\mathbf{SR_{12}}$ = “In the event of a physical server failure, the data:”.

Given that each criterion SR_j is further divided into an additional group of sub-criteria SR_{jk} with the following cardinalities: $m_{41} = 4$, $m_{42} = 2$, $m_{43} = 6$, $m_{44} = 2$, $m_{45} = 2$, $m_{46} = 2$, $m_{47} = 2$, $m_{48} = 2$, $m_{49} = 3$, $m_{410} = 4$, $m_{411} = 3$, $m_{412} = 3$, Formula 2 takes the following form:

$$\mathbf{SR} = \left\{ \bigcup_{j=1}^{12} \mathbf{SR_j} \right\} = \left\{ \bigcup_{j=1}^{12} \left\{ \bigcup_{k=1}^{m_{4j}} SR_{jk} \right\} \right\} = \tag{3}$$

$$\{\{SR_{11}, SR_{12}, SR_{13}, SR_{14}\}, \{SR_{21}, SR_{22}\},$$
$$\{SR_{31}, SR_{32}, SR_{33}, SR_{34}, SR_{35}, SR_{36}\}, \{SR_{41}, SR_{42}\}, \{SR_{51}, SR_{52}\},$$
$$\{SR_{61}, SR_{62}\}, \{SR_{71}, SR_{72}\}, \{SR_{81}, SR_{82}\}, \{SR_{91}, SR_{92}, SR_{93}\},$$
$$\{SR_{101}, SR_{102}, SR_{103}, SR_{104}\}, \{SR_{111}, SR_{112}, SR_{113}\}, \{SR_{121}, SR_{122}, SR_{123}\}\},$$

where: SR_{11} = “Restricted only to the cloud service’s own employees”; SR_{12} = “Access is granted to the cloud service’s employees and to the data center’s staff”; SR_{13} = “Access is granted to the cloud service’s employees, the cloud service’s customers, and the data center’s staff”; SR_{14} = “Access to the servers is denied to everyone except the data center’s staff”; SR_{21} = “Yes”; SR_{22} = “No”; SR_{31} = “key cards and a physical access key”; SR_{32} = “key cards”; SR_{33} = “by handing over a physical key”;

SR_{34} = “by individual paper passes”; SR_{35} = “by acquaintance (informally)”; SR_{36} = “not controlled in any way”; SR_{41} = “Yes, has autonomous power sources (generators, own substations, etc.)”; SR_{42} = “No”; SR_{51} = “Yes”; SR_{52} = “No”; SR_{61} = “Yes”; SR_{62} = “No”; SR_{71} = “Yes”; SR_{72} = “No”; SR_{81} = “Yes”; SR_{82} = “No”; SR_{91} = “Yes”; SR_{92} = “Partially, only conventional network firewalls are used (iptables, local FW, etc.)”; SR_{93} = “No”; SR_{101} = “Yes”; SR_{102} = “No, only IPS”; SR_{103} = “No, only IDS”; SR_{104} = “No IPS/IDS systems are present”; SR_{111} = “Completely impossible”; SR_{112} = “Possible upon prior request from the cloud service, with temporary access granted”; SR_{113} = “Anyone is able to connect to the servers after entering the server room”; SR_{121} = “Automatically fails over to an adjacent redundant server, and business procedures of the cloud service continue to run”; SR_{122} = “Are manually migrated to adjacent servers to restore business operability”; SR_{123} = “Are not preserved and are not restored”.

In the presented structure, criteria SR_1, SR_3, SR_9, SR_{10}, SR_{11} and SR_{12} offer more than a simple “Yes/No” answer, which allows the model to distinguish, within a single formal criterion, between a fully compliant, a partially compliant, and a non-compliant configuration – something that is not possible with the classic checklists typical of existing frameworks [3]-[7]. In particular, criterion SR_3 covers the full range of physical access controls observed in practice, while criterion SR_{12} formalizes the operational consequence of an equipment failure, rather than merely the formal existence of a backup policy on paper.

The SR module has a built-in point-based assessment system: each answer option is assigned a fixed score in the range from 0 to 5, where 5 corresponds to the best level of protection for the given criterion (for example, SR_{14} – access denied to everyone except data center staff, or SR_{121} – automatic failover with preservation of business processes), and 0 corresponds to the worst (for example, SR_{36} – access is not controlled at all, or SR_{123} – data is not preserved and is not restored). The resulting sum of scores across all twelve criteria and their sub-criteria forms the quantitative value of the SR parameter, which, together with the other parameter values of the CSP tuple, is used to compute the integral cyber-defense coefficient of the assessed cloud service in accordance with the base method [8].

4. Conclusions

In this research, the structure and the set-theoretic mathematical model of the server hardware module (parameter SR) were formalized as the fourth component of the developed generalized model for assessing the cyber defense state of cloud services. Twelve basic criteria were defined, covering physical access control, video surveillance, power-supply redundancy, climate control, fire suppression, intrusion detection at the physical and network perimeter, equipment lifecycle management, and equipment-failure behavior, together with their nested sub-criteria and answer options. The resulting description of the criteria closes an identified gap in existing physical-security frameworks, which offer qualitative checklists but lack a quantitative, per-criterion assessment mechanism comparable across cloud service providers. Further research is aimed at applying the developed point-based assessment method to the SR module on the example of a real

cloud provider, determining weighting coefficients for individual criteria, and integrating the SR module into the computation of the integral cyber-defense coefficient of a cloud service.

References

[1] Datadog. State of Cloud Security, 2025. URL: https://www.datadoghq.com/state-of-cloud-security/.

[2] CSA. Security Guidance for Cloud Computing v5, 2025. URL: https://cloudsecurityalliance.org/artifacts/security-guidance-v5.

[3] Uptime Institute. Data Center Site Infrastructure Tier Standard: Topology, 2026. URL: https://uptimeinstitute.com/tiers.

[4] Telecommunications Industry Association. ANSI/TIA-942-C Telecommunications Infrastructure Standard for Data Centers, 2025. URL: https://tiaonline.org/products-and-services/tia942certification/ansi-tia-942-standard/.

[5] International Organization for Standardization. ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection – Information security management systems – Requirements, Annex A.7 Physical controls, 2022. URL: https://www.iso.org/ru/standard/27001.

[6] National Institute of Standards and Technology. NIST Special Publication 800-53 Rev. 5, Security and Privacy Controls for Information Systems and Organizations, Physical and Environmental Protection (PE) family, 2020. URL: https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final.

[7] PCI Security Standards Council. Payment Card Industry Data Security Standard (PCI DSS) v4.0, Requirement 9: Restrict Physical Access to Cardholder Data, 2022. URL: https://www.pcisecuritystandards.org/standards/pci-dss/.

[8] Y.M. Pedchenko. Methods and models of assessing the state of cybersecurity of cloud services of information infrastructure objects : Doctor of Philosophy in Cybersecurity. Kyiv, 2025. 254 pp. URL: https://vchenarada.nau.edu.ua/wp-content/uploads/2025/05/Disertatsiya_Pedchenko_YEM_125_Kiberbezpeka.pdf.

[9] Y.V. Ivanchenko et al. The mathematical method for assessing the cybersecurity state of cloud services. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Srodowiska, 2025, 15(4), 137-141. URL: https://doi.org/10.35784/iapgos.7555.

[10] Y.M. Pechenko et al. Mathematical model of security cloud services assessment. Problems of scientific, technical and legal support for cybersecurity in the modern world : monograph / ed. by S. Semenov, M. Muchacki, Krakow. 2024. P. 13-21. doi: 10.24917/9788668020861

Autonomous Defensive Agents for Continuous Detection of AI-Powered Attacks on Educational Systems

Stepan Prokipchyn[1]

[1] *State University of Information and Communication Technologies, Solomianska 7, 03110 Kyiv, Ukraine*

Abstract

Agentic large language models have collapsed the time an attacker needs to convert a single foothold – a leaked credential, a known unpatched dependency, an undocumented entry point – into a full system compromise. What previously required weeks of manual reconnaissance can now be accomplished in minutes by parallel subagents that probe, adapt, and expand the attack surface autonomously. Educational systems are among the possible targets for such attacks because they concentrate personal data, corporate training material, and billing flows in systems typically maintained by small teams without dedicated security operations. This paper explores how threats from offensive agents can be answered symmetrically, and suggests two complementary defensive mechanisms: an agent that reviews the codebase and infrastructure configuration for vulnerabilities in a production-like sandbox; and an autonomous agent that analyses application, database, WAF, authentication, network, and cloud audit records. The contribution is the closed loop between the two.

Keywords

agentic AI, autonomous security, AI-enabled hacking, monitoring, vulnerability remediation

1. Statement of the problem

Offensive use of AI stopped being a projection in recent years. Anthropic detected a cyber campaign in which an agentic coding tool orchestrated through Model Context Protocol servers performed an estimated 80–90 % of the intrusion work against multiple organisations, with human operators intervening only at decision gates [1]. The UK AI Security Institute measured the time horizon for agent-driven cyber tasks that can be completed reliably and found it doubling every 4.7 months since late 2024, an acceleration from its own eight-month estimate published only three months earlier [2].

In June 2026 the Five Eyes cybersecurity agencies issued a joint advisory stating that the existence of AI systems prepared to overwhelm government and enterprise defences is “months, not years” [3]. An independent risk index that grades AI harms classifies AI-enabled hacking at crisis level with a worsening trend and confirmed evidence [4]. That qualification is the premise of this paper.

Two properties of this shift matter more to defenders than capability itself. The first is **parallelism** – a human attacker explores one hypothesis at a time, while an agentic system spawns dozens of subagents that probe multiple approaches concurrently. What's more, all these subagents can run on the same machine the attacker uses. The second is **compounding** – each successful step enlarges the reachable surface – a read-only database credential yields a configuration file, which yields a service token and so on. The interval between foothold and material impact, historically the defender's principal resource, contracts from weeks to hours.

Critically, this does not depend on frontier models. ESET's analysis of PromptLock, the first documented AI-powered ransomware, showed it generating cross-platform Lua payloads at

Corresponding author

s.prokipchyn@stud.duikt.edu.ua (S. Prokipchyn)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

runtime from a locally hosted open-weight 20-billion-parameter model served through Ollama [5]. Capability sufficient for opportunistic intrusion is therefore commodity, unmetered and unattributable – a different problem from the frontier-lab scenarios that dominate policy discussion.

Educational systems are exposed on five distinct axes:

- **Personal data.** Learner records contain Personally Identifiable Information, contact information and academic performance, often for minors.
- **Social engineering.** Stolen course progress data allows phishing against students and instructors: a message naming a real course, deadline, and instructor can be enough to deceive.
- **Corporate confidentiality.** The platform may deliver internal corporate materials.
- **Billing abuse.** Payment-processing paths are directly monetisable, and abuse there may not surface as a security alert at all.
- **Availability and ransom.** Assessment windows and enrolment periods are hard deadlines, raising the coercive value of encryption or denial of service far above the system's ordinary criticality.

Our earlier analysis of this class of system distinguished three layers at which AI is introduced – internal automation agents, AI-assisted engineering, and user-facing AI features – and showed that each layer elevates pre-existing access-control weaknesses rather than creating new ones [6]. The present work takes that risk decomposition as its starting point and asks the converse question: how the same agentic capability can be turned to the defence of the platform.

These systems are typically maintained by a few engineers with no dedicated security function. The defensive question is therefore not which controls are theoretically correct, but which a small team can actually operate against an adversary.

2. Scientific novelty

Autonomous vulnerability discovery and LLM-based log analysis have both been demonstrated in isolation. At DARPA's AI Cyber Challenge, concluded in August 2025, several cyber reasoning systems ran autonomously across dozens of challenge projects, discovering 18 previously unknown real-world vulnerabilities and patching them in about 45 minutes on average [7, 8]. This speed is unattainable for human teams. Google's Big Sleep agent identified an existing, previously unknown vulnerability in SQLite, the first widely reported case of an AI agent pre-empting a live threat actor [9]. On the defensive operations side, agent security taxonomies [10], the OWASP Top 10 for LLM Applications [11], MITRE ATLAS [12], and layered security systems [13] describe the risk surface of agentic systems.

What is not established is the **operational coupling** of these capabilities under the constraints of a small product team, nor their specialisation to the educational domain. Existing work optimises either for competition conditions – a bounded corpus, unlimited compute, no production system to avoid breaking – or for enterprise SOCs with staffed escalation tiers. The novelty claimed here is threefold:

1. **A closed feedback loop between proactive remediation and production monitoring.** Anomalies detected in production are converted into targeted audit tasks for the remediation agent. In turn, the exploitability evidence generated by the remediation agent

is used to create detection signatures and expected-behaviour baselines for the monitoring agent. This guarantees that even if fixes proposed by the first agent are rejected by the human engineer, the second agent knows about the possible attack vector and monitors for it. Neither component is novel on its own; the feedback loop between them is what transforms two separate tools into a unified control loop.

2. **Exploitability-gated remediation instead of version-gated remediation.** The prevailing practice – bump every flagged dependency to the latest – optimises a scanner metric rather than risk, and pays for it in breaking changes and in newly introduced vulnerabilities in the target version. Reachability and sandbox-verified exploitability are substituted as the gating criterion, with regression scope analysed and tested before the change is proposed.
3. **Cost-tiered escalation as a design constraint, not an optimisation.** Monitoring full log volume with a frontier model is economically impossible for this class of system. Treating model selection as a first-class architectural parameter – a cheap classifier at ingest, an expensive reasoner on escalation – is what makes continuous coverage feasible at all.

3. Proposed solution

3.1. Proactive vulnerability detection and remediation

The first component is a separate AI agent that has access to the system codebase and operates in an isolated environment provisioned to imitate production, but seeded exclusively with synthetic data. It runs on a schedule and on trigger, and it has no credentials to production.

The dependency case illustrates the pipeline most clearly, because it is where the gap between current practice and achievable practice is widest:

1. **Detection.** A known vulnerability is identified in a transitive or direct dependency from advisory feeds correlated against the resolved dependency graph.
2. **Exploitability analysis.** The agent determines whether the vulnerable code path is reachable from the application's entry points at all, and under what preconditions – authentication state, feature flag, configuration value, input shape. A large share of advisories are unreachable in a given application, and distinguishing these is the difference between a queue that gets worked and a queue that gets ignored.
3. **Sandbox exploitation testing.** When an exploitable path is possible, the agent constructs a test scenario and executes it in the sandbox. This is an important gate, because the next three phases are expensive: analysing the entire codebase for possible regressions, fixing all the places and preparing final change request should be avoided for something that cannot be reproduced. A successful reproduction turns hypothesis into a confirmed vulnerability. If reproduction fails, the outcome and the reasoning behind it are recorded rather than silently discarded.
4. **Regression scope analysis.** Before posting an upgrade, the agent reads release notes and changelogs of the candidate version, examines the changes against actual call sites in the codebase, and enumerates the behaviours at risk. It also checks whether the candidate version carries advisories of its own.
5. **Test synthesis and verification.** Behaviours identified as at risk that lack coverage are covered with generated tests, which must pass against the pre-upgrade version before they are trusted as regression detectors. The upgrade is then applied in the sandbox and the suite executed against the running system.

6. **Evidence-bearing change request.** The change is submitted with the finding itself, the proof-of-concept scenario, the regression analysis, the generated tests, and the sandbox run results attached. A human engineer approves or rejects.

The final step is not a limitation to be engineered away. Autonomous patching at competition scale still produces semantically incorrect patches at a substantial rate even when they pass automated validation [8], and the agent's own privileges are themselves an attack surface – credential exposure, over-broad scoping, and unintended write actions are amplified rather than created by agent autonomy [6]. Human approval keeps the defensive agent from becoming the most efficient attack path into the system it defends. The pipeline generalises beyond dependencies to application code, configuration drift, IAM over-provisioning, and exposed storage.

3.2. Persistent monitoring of the production environment

The second component addresses the real current state of the production system. It allows determining if an attacker is already inside, or exploiting something not yet known. It ingests and correlates:

- WAF and edge access logs (probing patterns, parameter fuzzing, enumeration) are the richest sources of information for the majority of application-level attacks;
- application logs (errors, latency, unusual code paths) add domain meaning to raw access logs;
- database change logs (bulk reads, mass updates) reveal the actual range of data under attack, especially in data-harvesting attacks;
- authentication logs can uncover the entry point of the attack, as most application-level attacks start with leaked credentials or other types of privileged account takeover;
- network flow logs and cloud audit records can be useful for attacks on the infrastructure.

The proposed architecture is a two-layered system. An inexpensive, fast model performs first-pass triage, scoring combined logs for anomalies. Its task is deliberately narrow – decide whether an event window deserves attention – which is what permits it to run over full log volume.

In the event of escalation, a full-capability model performs a second pass over the same event window. It reconstructs a hypothesis of what is happening, assigns a confidence and a severity, and either dismisses with a recorded rationale or notifies the responsible engineers. This stronger model can have access to a project-specific knowledge base of possible weaknesses of the system, generated by the remediation agent. This makes it more powerful compared to a raw LLM.

The more data sources we combine, the better the reasoning will be. The context window of modern models can fit up to one million tokens, which is more than enough to fit the knowledge base and a significant timeframe of converged data. This allows to detect a failed login from a new location, followed by a token mint, a read of a table that service has never touched, and egress to an unfamiliar destination – the kind of cross-source story that costs a human analyst hours to assemble and an agent seconds.

Such deep analysis can take minutes to complete, which is still comparable with the attack pace. Alert latency becomes important: first-pass triage must run either on the event stream or on periodic small batch aggregation, accepting a higher false-positive rate at the cheap tier in exchange for the escalation tier's ability to dismiss it cheaply.

3.3. Safety requirements on the defensive agents

Both components are themselves agentic systems operating inside the security perimeter, and inherit the corresponding risks, and fall under the mitigations we previously argued for agents acting inside educational systems: least-privilege scoping, short-lived credentials, an audit layer, and human-in-the-loop approval above a risk threshold [6, 10, 11]. Several additional security measures are required to avoid introducing any additional security risks. The monitoring agent is read-only over log sinks. Some production log records may require prior sanitisation to prevent the exposure of sensitive data. Auxiliary tooling can be used to analyse and review the agent's actions.

4. Conclusions

Offensive AI agents allow decreasing the interval between a single entry point and actual compromise from weeks to hours or even less. E-learning platforms and other educational systems are at risk and often lack sufficient resources to invest in cybersecurity. Small teams often focus on new features in a competitive market and don't have the expertise needed for proper SecOps.

The paper describes an architecture in which two AI agents are combined into a feedback loop. An autonomous remediation agent analyses source code and configuration, reproduces potential attacks in an isolated environment and prepares evidence-bearing changes for human approval. This agent can also build a live knowledge base of project-specific intrusions.

The second agent monitors production telemetry and detects real-time anomalies. It uses the first agent's knowledge base to augment its training data. Any real incident gets written down and becomes a new investigation vector for the first agent.

The result is the closed loop between them – production anomalies become audit targets, remediation evidence becomes monitoring baselines.

Both components are described here at the architectural level. Real-world testing on a production educational platform – measuring detection latency and correctness at each tier, remediation acceptance rate under human review, and total operating cost per protected service – is ongoing work.

References

[1] Anthropic, Disrupting the first reported AI-orchestrated cyber espionage campaign, Anthropic Threat Intelligence Report, 2025. URL:.

[2] UK AI Security Institute, How fast is autonomous AI cyber capability advancing?, 2026. URL: https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing.

[3] Australian Signals Directorate ACSC, Canadian Centre for Cyber Security, New Zealand NCSC, UK NCSC, US CISA, Joint advisory on AI-enabled cyber threats, 2026. URL: https://www.licentium.io/post/five-eyes-agencies-joint-ai-cyber-threat-statement-june-2026

[4] The Multiverse School, AI Problems Index – AI Risk Atlas: AI-enabled hacking, 2026. URL: https://themultiverse.school/x/ai-problems-index.

[5] ESET Research, First known AI-powered ransomware uncovered: PromptLock, WeLiveSecurity, 2025. URL: https://www.welivesecurity.com/en/ransomware/first-known-ai-powered-ransomware-uncovered-eset-research/.

[6] S. Prokipchyn, Mitigating AI-driven security risks in educational software systems, in: ITSec 2026: Proceedings of the XV International Scientific and Technical Conference, TNTU–SUICT, Ternopil–Kyiv, 2026, pp. 286–287. URL: https://drive.google.com/file/d/1Z_ufhNPv6XzsPqkaPoBoz9hk1lDFoFcP/view

[7] DARPA, AI Cyber Challenge (AIxCC): final competition results, DEF CON 33, 2025.
[8] SoK: DARPA's AI Cyber Challenge (AIxCC), arXiv:2602.07666, 2026. URL: https://www.usenix.org/system/files/usenixsecurity26-zhang-cen.pdf
[9] Google, Big Sleep: AI agent discovery of CVE-2025-6965 in SQLite, Google Security Blog, 2025. URL: https://blog.google/innovation-and-ai/technology/safety-security/cybersecurity-updates-summer-2025/
[10] Y. Tang, et al., Security of LLM-based agents regarding attacks and defenses, Information Fusion, 2025. doi:10.1016/j.inffus.2025.103941. URL: https://www.researchgate.net/publication/397754198_Security_of_LLM-based_Agents_Regarding_Attacks_Defenses_and_Applications_A_Comprehensive_Survey
[11] OWASP Gen AI Security Project, OWASP Top 10 for LLM Applications 2025, 2025. URL: https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025.
[12] MITRE, ATLAS: Adversarial Threat Landscape for AI Systems. URL: https://atlas.mitre.org.
[13] Y. Gao, S. Wu, A four-layer security governance framework for LLM-based AI agents, Journal of Artificial Intelligence Practice, 2025. doi:10.23977/jaip.2025.080406. URL: https://www.clausiuspress.com/article/17139.html

Cognitive Correlation of Cyber Threats across OSI Model Layers for Incident Response Support

Dmytro Prokopovych-Tkachenko[1,2,3]

[1] *University of Customs and Finance*
[2] *State Scientific Institution "Institute of Information, Security and Law of the National Academy of Legal Sciences of Ukraine"*
[3] *State University of Telecommunications*

Abstract

Security sensors commonly observe selected protocol layers, leaving multi-stage incidents fragmented into unrelated alerts. This paper proposes a cognitive correlation model that joins telemetry mapped to the seven Open Systems Interconnection layers with asset context, event provenance, and adversary-behavior hypotheses. Its architecture includes layer-aware collectors, schema normalization, clock-quality control, a provenance graph, hypothesis competition, risk contextualization, and analyst feedback. The OSI model is used only as an observation coordinate; tactics, assets, identities, and causal relations remain separate. Proposed emulation scenarios cover address-resolution manipulation, session interception, routing and name-service anomalies, credential replay, and application-layer injection. Planned metrics include precision, recall, F1, false-alert rate, detection latency, layer-localization accuracy, calibration, explanation sufficiency, and overhead. All scenario values are illustrative; no measured performance is claimed. The expected contribution is a reproducible, explainable method for linking weak cross-layer signals while keeping critical incident-response actions under authorized human control.

Keywords

cognitive cybersecurity, OSI model, alert correlation, knowledge graph, incident response, explainable security analytics, security operations center

1. Introduction and Problem Statement

Communication systems expose heterogeneous telemetry: interface state, address-resolution records, routing flows, transport changes, sessions, identities, and application logs. A single intrusion may alter several planes without yielding a decisive indicator at one sensor. Each event can be benign alone, while a plausible sequence around one asset may reveal an incident. Monitoring tools preserve alerts but lose these relationships. The Open Systems Interconnection model supplies a common vocabulary for communication functions [1]. It is not a threat model; here, its layers identify where evidence originated.

Because stacks do not map neatly to OSI boundaries, layer attribution remains uncertain and separate from tactics, assets, identities, vulnerabilities, and impact. Incident response requires detection, prioritization, analysis, containment, recovery, and learning [2]. Analysts must decide whether events concern the same entity, timestamps are trustworthy, and upstream changes explain downstream symptoms. This paper asks how observations from different OSI layers can become explainable incident hypotheses without presenting heuristic scores as attribution. It proposes competing hypotheses, explicit supporting and conflicting evidence, provenance metadata, and versioned analyst correction. The outcome is an inspectable pipeline for evaluation on reproducible, authorized scenarios before deployment.

Corresponding author
omega2417@umsf.edu.ua (D. Prokopovych-Tkachenko)
0000-0002-6590-3898 (D. Prokopovych-Tkachenko)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Related Work and Scientific Novelty

Layered reasoning has two distinct roles in security analysis. The OSI model standardizes interconnection functions and provides a stable vocabulary for locating observations [1]. It does not define adversary behavior, and practical TCP/IP implementations often blur its boundaries. Classic protocol analysis demonstrates that weaknesses can span addressing, routing, authentication, and trust assumptions [8]. Consequently, this work treats a candidate layer as uncertain observation metadata rather than an attack label. Threat intent, identity, asset criticality, dependency, and operational impact are represented on separate axes. Existing operational representations address adjacent parts of the problem.

STIX defines structured cyber-threat objects and relationships [3], ATT&CK organizes documented adversary behaviors [4], and NIST incident-response guidance links detection with preparation, response, recovery, and improvement [2]. PROV-O supplies reusable concepts for recording entities, activities, agents, and derivation [7]. These sources support interoperability and traceability, but they do not prescribe how a SOC should combine temporally uncertain evidence from several protocol layers, maintain competing explanations, or expose contradictions in a calibrated decision process. Alert clustering can reduce duplication and support root-cause investigation [6].

Machine learning may discover additional relations, yet operational network detection differs from closed-world classification: labels are incomplete, traffic changes, and a topology can leak scenario-specific shortcuts [5]. A cross-layer model trained and tested on randomly mixed records could therefore appear accurate while learning sensor placement rather than incident structure. The proposed protocol separates complete scenario runs, holds out topology variants, and treats learned similarity as one evidence source among provenance, temporal, semantic, and graph constraints. The explicit scientific novelty has four components. First, every event receives dual coordinates: a sensor-and-layer observation coordinate and a threat coordinate linking behaviors, entities, assets, and dependencies.

Second, the graph retains parser status, clock quality, transformation provenance, missing observations, and conflicting evidence instead of collapsing them into an opaque vector. Third, competing hypotheses combine temporal, topological, identity, and semantic relations while preserving an inspectable evidence path for each recommendation. Fourth, analyst feedback can revise bounded weights or rules only through a versioned audit trail, preventing silent rewriting of earlier cases. The contribution is a reproducible representation and decision-support procedure, not a universal detector or attribution mechanism. Proposed evaluation compares independent alerts, temporal rules, simpler graph correlation, and the full model.

It tests incident linkage, layer localization, score calibration, explanation traceability, latency, and resource overhead. These are prospective, falsifiable criteria; no performance advantage or completed experiment is claimed.

2.1. Formal Model

$$q_e = q_src * q_time * q_parse * \exp(-lambda * dt_e) \quad (1)$$

where q_e: event confidence; q_src, q_time, q_parse: quality factors; dt_e: nonnegative event age; lambda: proposed decay rate. Weights evidence by provenance quality and freshness.

$$S_h = b_h + \sum_e(q_e * phi_e) + \sum_ij(psi_ij) - beta * C_h \quad (2)$$

where S_h: hypothesis support; b_h: prior; phi_e: event compatibility; psi_ij: relation compatibility; C_h: contradiction score; beta: penalty. Combines graph evidence while penalizing explicit contradictions.

$$U = -\sum_h(p_h * \log(p_h)) / \log(N_h) \quad (3)$$

where U: normalized ambiguity; p_h: normalized relative hypothesis support; N_h: number of active hypotheses, with N_h greater than one. Escalates ambiguous cases instead of asserting attribution.

3. Proposed Solution and Method

Read-only collectors acquire interface, address-resolution, flow, transport, authentication, and application events. Each event envelope records event and receipt times, source, parser version, entity identifiers, candidate OSI layer, raw-record hash, and quality fields. Normalization aligns identifiers while retaining originals. Clock uncertainty lowers confidence rather than deleting evidence; sensitive fields are tokenized. A property graph links events to sensors, assets, identities, services, locations, vulnerabilities, and behavior hypotheses. Temporal edges encode ordering; topology edges encode communication and dependency; provenance edges encode generation and derivation.

STIX and ATT&CK mappings remain separate from OSI attribution, avoiding false equivalence between a protocol layer and attacker behavior. The engine generates candidates from graph motifs that match ordered behavior templates. It combines source quality, time decay, semantic compatibility, cross-layer relations, and conflict penalties, then contextualizes hypotheses with asset criticality. The equations define event confidence, hypothesis support, and ambiguity. Scores have no probabilistic meaning unless held-out calibration demonstrates it. Each explanation lists supporting and conflicting paths, missing observations, evidence sensitivity, and a proposed playbook. Analysts may confirm, reject, merge, or defer hypotheses.

Feedback is versioned and changes bounded weights only after governance review. Isolation, credential revocation, and route modification require an authorized person or separately approved policy. The architecture supports traceability from raw evidence to recommendation and enables ablation of provenance, cross-layer relations, context, and feedback.

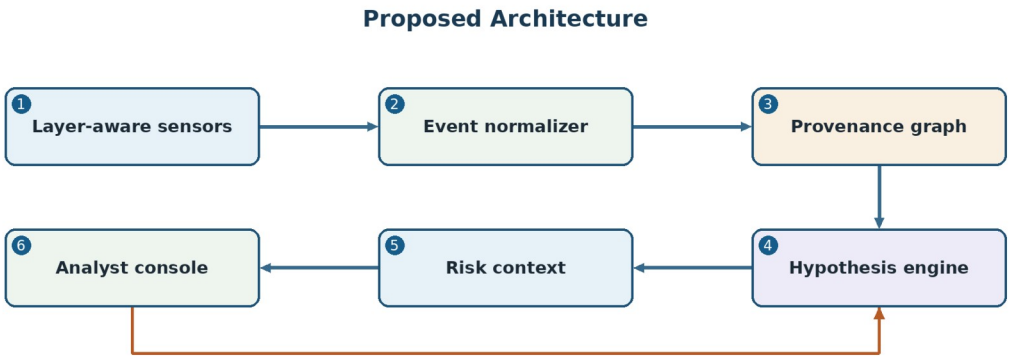


Figure 1: Proposed provenance-preserving cross-layer correlation architecture

4. Evaluation Design

Evaluation is proposed on an isolated enterprise emulator with endpoints, switching, routing, identity, name services, web/API services, and centralized logs. Authorized scenarios cover benign changes, address-resolution manipulation, session interception, routing and name-service anomalies, credential replay, and application injection. Table values are illustrative, not observations. Complete scenario runs, not records, form development, calibration, and held-out

partitions; topology variants test transfer. Baselines include independent alerts, temporal rules, graph correlation without hypothesis competition, and the full model. Ablations remove provenance, cross-layer terms, asset context, or feedback.

Metrics include precision, recall, F1, false alerts, linkage accuracy, latency, layer localization, Brier score after calibration, explanation traceability, throughput, and memory. Paired bootstrap comparisons report confidence intervals and effect sizes. Failure analysis separates sensing, identity-resolution, timing, mapping, and ranking errors. No empirical claim precedes protocol execution.

Table 1
Illustrative proposed validation matrix (design values only; no empirical results)

Proposed scenario	Cross-layer evidence	Illustrative design	Primary metric	Expected check
Benign mixed workload	L2 state; L3 flows; L7 logs	Illustrative: 12 × 30-minute runs	False alerts/hour	Benign changes remain distinct
Address-resolution manipulation	L2 mapping; L3 relation; L5 session	Illustrative: 10 × 30-minute runs	Incident-linkage F1	No single-event overclaim
Routing anomaly	L3 route; L4 reset; L7 error	Illustrative: 3 topologies; 24 runs	Layer-localization accuracy	Upstream cause remains visible
Name-service manipulation	L3 path; L5 query; L7 mismatch	Illustrative: 12 cases/topology	Detection latency	Ambiguity raises uncertainty
Credential replay	L5 session; L6 context; L7 identity	Illustrative: 4 contexts; 40 runs	Calibrated Brier score	Risk reflects context
Application injection; conflicting sensors	L4 flow; L6 encoding; L7 audit	Illustrative: 20 cases; 50% conflicts	Conflict-disclosure score	Conflicts remain visible

Note. Values are illustrative design parameters, not measured results.

5. Discussion

The model separates evidence location from threat interpretation: OSI layers describe where an observation arose, while behavior, identity, asset, topology, and provenance relations explain why observations may belong to one incident. Competing hypotheses, explicit contradictions, and ambiguity scores keep uncertainty visible. Provenance factors also prevent duplicated sensor outputs from appearing as independent corroboration. Important limitations remain. Encrypted, virtualized, software-defined, and cloud-native environments may provide only indirect layer evidence. Clock correction, entity resolution, sensor placement, ATT&CK mapping, and analyst feedback can introduce false links or bias.

A laboratory emulator cannot reproduce production diversity or adaptive adversaries, so access control, protected evidence storage, versioning, and audited overrides are required before deployment. The central conclusion is that cross-layer correlation is scientifically defensible only when observation location, threat semantics, provenance, and uncertainty remain independently inspectable.

6. Conclusions

The contribution is a dual-coordinate event model, provenance graph, competing-hypothesis mechanism, and human-governed validation protocol; it is not validated operational superiority. Future work should implement the emulator, preregister parameters, compare baselines and ablations, and report negative or inconclusive cases. Tests should cover missing telemetry, compromised sensors, topology change, manipulated graph evidence, explanation quality, and

analyst cognitive burden. Poor calibration, unstable transfer, excessive overhead, or ambiguous explanations should trigger revision rather than automation.

Declaration on Generative AI

Generative AI tools were employed exclusively for language editing, grammatical refinement, and visual figure generation. All scientific content, analysis, and conclusions were authored, verified, and finalized by the author(s).

References

[1] ISO/IEC 7498-1:1994, Information technology — Open Systems Interconnection — Basic Reference Model: The Basic Model. International Organization for Standardization, Geneva, 1994. https://www.iso.org/standard/20269.html

[2] A. Nelson, S. Rekhi, M. Souppaya, and K. Scarfone, Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile, NIST Special Publication 800-61 Rev. 3, April 2025. https://doi.org/10.6028/NIST.SP.800-61r3

[3] B. Jordan, R. Piazza, and T. Darley, eds., STIX Version 2.1, OASIS Standard, 10 June 2021. https://docs.oasis-open.org/cti/stix/v2.1/os/stix-v2.1-os.html

[4] MITRE, MITRE ATT&CK: Enterprise Matrix, authoritative online knowledge base. https://attack.mitre.org/matrices/enterprise/ (accessed 29 August 2026)

[5] R. Sommer and V. Paxson, Outside the Closed World: On Using Machine Learning for Network Intrusion Detection, in 2010 IEEE Symposium on Security and Privacy, Berkeley/Oakland, CA, USA, 2010, pp. 305–316. https://doi.org/10.1109/SP.2010.25

[6] K. Julisch, Clustering Intrusion Detection Alarms to Support Root Cause Analysis, ACM Transactions on Information and System Security, vol. 6, no. 4, 2003, pp. 443–471. https://doi.org/10.1145/950191.950192

[7] T. Lebo, S. Sahoo, and D. McGuinness, eds., PROV-O: The PROV Ontology, W3C Recommendation, 30 April 2013. https://www.w3.org/TR/prov-o/

[8] S. M. Bellovin, Security Problems in the TCP/IP Protocol Suite, ACM SIGCOMM Computer Communication Review, vol. 19, no. 2, 1989, pp. 32–48. https://doi.org/10.1145/378444.378449

An Integrated Decision-Support Loop for Cyber-Incident Neutralization in Military Command and Control: From the Evaluation Image to an Admissible Response

Hennadii Rybachok[1]

[1] *National Defence University of Ukraine, Povitroflotskyi Ave. 28, 03049 Kyiv, Ukraine*

Abstract

For military command and control (C2) bodies of the Armed Forces of Ukraine a cyber incident is a command problem: the response option has to be chosen under time pressure, and its justification must survive an incident review. This contribution presents the architecture of an information technology that supports such decisions as a loop of seven stages: evaluation of the incident, assembly of the decision context, setting of the target parameters (the required level of neutralization and the admissible residual risk), formation of the initial set of response alternatives by a response knowledge base, admissibility checking with criteria-based ranking, prioritization across concurrent incidents, and the decision package for the C2 official. The novelty lies in the integration of previously published components into one loop with explicitly defined hand-off objects between the stages and with an explanation carried through to the official. The loop is illustrated by a case of a phishing campaign with privilege escalation in a planning segment.

Keywords

decision support, cyber incident, neutralization, military command and control, explainability

1. Problem statement

Cyber incidents against the information and communication systems of the Armed Forces of Ukraine have become a routine load on military command and control (C2) bodies: denial-of-service attacks on communication nodes, phishing that ends in privilege escalation, and injections into the databases of support subsystems arrive during the active phases of operations, when the decision time is measured in minutes and the duty shift is already occupied with the command cycle itself. In this setting the response is a command decision. A shutdown of a node, a switchover to a standby channel, or a delay of a configuration rollback changes the ability to command troops, and an incorrect choice costs more than the affected server.

The regulatory and process frameworks available to the duty shift describe how the response should be organized, not what a sufficient response is for a particular incident in a particular command situation. ISO/IEC 27035-1 fixes the phases of incident management [1]; NIST SP 800-61r3 embeds incident handling into the cybersecurity risk management cycle [2]; the Law of Ukraine "On the Basic Principles of Ensuring Cybersecurity of Ukraine" sets the limits of authority and the duties of information exchange [3]. None of these documents answers the questions that the duty officer actually faces: how significant this incident is for the mission, what level of neutralization is required and what residual risk can be accepted, which response options exist and which of them are admissible under the current mode, which option to recommend when effectiveness, time, and resources conflict, and how to justify the recommendation before the official who decides.

Corresponding author

ryba4okgen@gmail.com (H. Rybachok)

0009-0002-8513-5133 (H. Rybachok)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

These questions are answered today from personal experience, scattered instructions, and memory of past incidents. The result is well known from practice: comparable incidents are handled differently by different shifts, the justification of a decision cannot be reconstructed after the fact, and the experience of a shift leaves the C2 body together with the rotation. The problem addressed by the author's research is therefore not a single model but an information technology: a coherent sequence of formal transformations that takes the primary description of a cyber incident and produces a justified, admissible, and explainable recommendation for the C2 official, while leaving the decision itself to that official.

2. Scientific novelty versus existing works

Response automation has advanced mainly through playbooks and orchestration. The CACAO specification standardizes machine-readable security playbooks [4], and SOAR platforms increasingly use machine learning to enrich alerts and to trigger prepared scenarios [5]. Playbooks reproduce known action sequences well, yet they neither compute how much neutralization is enough for a given incident nor represent the admissibility constraints of a C2 body as a separate layer; the recommendation logic of learned models remains hidden from the person who has to defend the decision.

Formal countermeasure selection has been surveyed in depth: reaction frameworks compare countermeasures by cost, effectiveness, and time [6], and knowledge resources such as MITRE ATT&CK and D3FEND provide a vocabulary of adversary techniques and defensive countermeasures [7, 8]. These works supply the building blocks of a decision but do not define the target of the decision (the required level of neutralization and the admissible residual risk) in terms of the command situation, which is exactly where the military case differs from the corporate one.

Works on cyber defence of military information and communication systems concentrate on the detection side. A SIEM architecture with fuzzy inference for the databases of military systems has been proposed in [9]; multi-criteria and fuzzy schemes for ranking response options are also known [6]. The stage between detection and the choice of an alternative, where the incident is evaluated in mission terms and the space of admissible options is formed, has not been described as an integrated technology for a C2 body.

The author's earlier results cover individual links of this stage. The model of cyber-incident neutralization under time and resource constraints represents the response as a chain of information transformations from the incident description to the decision package [10]. The methodology of evaluating the effectiveness of neutralization defines the system of indicators and criteria and the evaluation image of the incident [11]. The factors and the context of the C2 decision have been structured as a package of input data with a status of what is known, assumed, and unknown [12]. What has not been shown before, and what constitutes the novelty of this contribution, is the integration of these links into one loop with explicitly defined hand-off objects between the stages: the evaluation image, the pair of target parameters, the initial set of alternatives, the admissible subset, the ranked recommendation, and the decision package. Each hand-off object is produced by one component and consumed by the next without re-evaluation, and the responsibility of every component is bounded so that no component makes the decision.

3. Proposed solution overview

The technology is organized as a loop of seven stages (Fig. 1), built on the neutralization model [10]. Its input is the primary description of an incident INC_i received from monitoring and reports; its output is the decision support package $Result_i$ that the official considers.

Stage 1, evaluation. The primary description is transformed into the evaluation image O_i^{eval} by the system of indicators and criteria of the effectiveness evaluation methodology [11]: the integral significance of the incident, the group profile of indicators, the military-management impact, the set of triggered critical features, and the quality of the information basis. The image is the only representation of the incident that the later stages use; they do not return to raw data.

Stage 2, context. The decision context is assembled as a package of factors, mission requirements, normative constraints, asset criticality, and available threat intelligence, with each element marked as known, assumed, or unknown [12]. The context is not part of the incident, and keeping it separate allows the same incident to be handled differently in different phases of an operation without changing the evaluation.

Stage 3, target setting. From the image and the context the technology computes the pair of target parameters: the required level of neutralization and the admissible residual risk. The pair is monotone with respect to the significance and the military-management impact of the incident, and it is set before any response option is considered. This is the point where "sufficient" is defined for the C2 body rather than assumed.

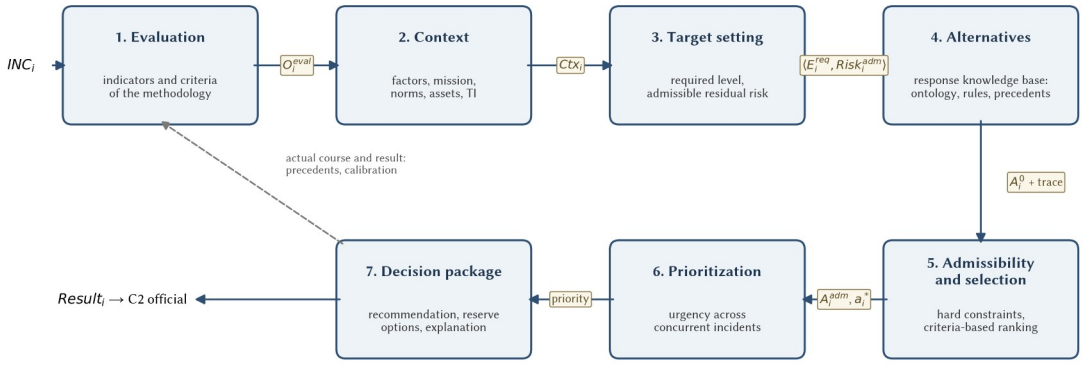


Figure 1: The seven-stage decision-support loop for cyber-incident neutralization (developed by the author)

Stage 4, formation of alternatives. A response knowledge base with an ontological, a production, and a precedent component returns the initial set of response alternatives, each accompanied by a record of where it came from: the ontological chain, the precedents used, and the rules that fired. Options prohibited by normative or mode rules are removed at this stage, and the removal is recorded with the rule and its source.

Stage 5, admissibility and selection. Each alternative is checked against the target pair, the time and resource limits, and the normative and mode predicates; the admissible alternatives are ranked by a criteria-based procedure that weighs the expected effectiveness, the residual risk, the time to effect, and the resource cost. Ranking is deliberately separated from admissibility: an option that violates a hard constraint is never traded off against effectiveness.

Stage 6, prioritization across incidents. When several incidents compete for the same personnel, channels, and time, the recommended options are ordered by an urgency index with protection against starvation of low-priority incidents, and the shared budget is allocated before the recommendations are issued.

Stage 7, decision package. The recommended alternative, the reserve options, the inference trace, and the expected values of the indicators form $Result_i$. The official approves, modifies, or rejects the

recommendation; the actual course and result of the response are recorded and, after validation, extend the precedent component and calibrate the indicators.

The stages, their hand-off objects, and the underlying components are summarized in Table 1.

Table 1
Stages of the loop, their hand-off objects, and underlying components

Stage	Input	Hand-off object produced	Underlying component
1. Evaluation	primary description INC_i	evaluation image O_i^{eval}	system of indicators and criteria [11]
2. Context	mission, norms, assets, threat intelligence	context package Ctx_i with known/assumed/unknown marks	taxonomy of decision factors [12]
3. Target setting	O_i^{eval}, Ctx_i	required level E_i^{req} and admissible residual risk $Risk_i^{\text{adm}}$	monotone target-setting rule
4. Alternatives	O_i^{eval}, Ctx_i, resources, norms, mode	initial set A_i^0 with inference trace	response knowledge base (ontology, rules, precedents)
5. Admissibility and selection	A_i^0, target pair, limits	admissible subset A_i^{adm} and recommended a_i^*	admissibility predicates, criteria-based ranking
6. Prioritization	recommendations of concurrent incidents	order and shared-budget allocation	urgency index with anti-starvation
7. Decision package	a_i^*, reserve options, traces	$Result_i$ for the C2 official; validated precedent	neutralization model [10]

Source: developed by the author.

Every hand-off object is a formal data structure with a defined producer and consumer, so the technology can be introduced incrementally: a C2 body may start with stages 1, 3, and 5 and add the knowledge base and cross-incident prioritization later. The explanation is not an add-on: the inference trace of stage 4, the admissibility record of stage 5, and the priorities of stage 6 are carried into the package.

An illustrative case shows the loop at work. A phishing campaign followed by privilege escalation is detected in a segment that supports the planning of an operation. Stage 1 assigns a high integral significance because the affected resource belongs to the command loop and the critical feature "access to planning data" is triggered. Stage 2 records the active phase of the operation, the combat mode, and the absence of redundancy for the segment. Stage 3 sets a critical required level with a low admissible residual risk. Stage 4 returns four candidate packages, two of them backed by precedents, and removes a full segment shutdown under the command continuity rule. Stage 5 confirms two admissible packages (credential revocation with session isolation, and isolation combined with a staged rebuild of the affected hosts) and ranks the first one higher on time to effect. Stage 6 orders this incident above a concurrent low-impact scanning event that competes for the same administrators. Stage 7 hands the official a package in which each of these steps is visible, and the record of the case becomes a precedent for the next shift.

4. Conclusions with main results

The contribution has presented the architecture of an information technology that supports the decisions of military C2 bodies on cyber-incident neutralization as a loop of seven stages with formally defined hand-off objects. The loop integrates the results published by the author earlier, namely the neutralization model [10], the effectiveness evaluation methodology [11], and the

structure of decision factors and context [12], and adds the components that form the target of the response, the space of admissible options, and the cross-incident priorities.

The main results are the following. The target of the response is computed from the evaluation image and the context before any option is considered, so "sufficient" is no longer an expert assumption. Admissibility is separated from ranking, so hard normative and mode constraints are never traded off against effectiveness. The explanation of the recommendation is carried through all stages to the official. The decision remains with the official, and the actual outcome of every case returns to the technology as a precedent and as calibration data.

Further work concerns the experimental assessment of the loop on the materials of real duty shifts and its integration with the monitoring and orchestration tools already deployed.

Declaration on Generative AI

During the preparation of this work, the author used AI-based tools in order to: Grammar and spelling check; Literature search, with subsequent verification of all sources against primary records; Generate the diagram in Figure 1 from the author's specification. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] International Organization for Standardization, ISO/IEC 27035-1:2023 Information technology — Information security incident management — Part 1: Principles and process, ISO, Geneva, 2023. URL: https://www.iso.org/standard/78973.html.

[2] A. Nelson, S. Rekhi, M. Souppaya, K. Scarfone, Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile, NIST Special Publication 800-61r3, National Institute of Standards and Technology, Gaithersburg, MD, 2025. doi:10.6028/NIST.SP.800-61r3.

[3] Verkhovna Rada of Ukraine, On the Basic Principles of Ensuring Cybersecurity of Ukraine, Law of Ukraine No. 2163-VIII, 2017. URL: https://zakon.rada.gov.ua/laws/show/2163-19 (in Ukrainian).

[4] OASIS Open, CACAO Security Playbooks Version 2.0, Committee Specification 01, OASIS, 2023. URL: https://docs.oasis-open.org/cacao/security-playbooks/v2.0/security-playbooks-v2.0.html.

[5] J. Kinyua, L. Awuah, AI/ML in security orchestration, automation and response: Future research directions, Intelligent Automation & Soft Computing 28 (2021) 527–545. doi:10.32604/iasc.2021.016240.

[6] P. Nespoli, D. Papamartzivanos, F. Gómez Mármol, G. Kambourakis, Optimal countermeasures selection against cyber attacks: A comprehensive survey on reaction frameworks, IEEE Communications Surveys & Tutorials 20 (2018) 1361–1396. doi:10.1109/COMST.2017.2781126.

[7] B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, C. B. Thomas, MITRE ATT&CK: Design and Philosophy, MITRE Product MP180360R1, The MITRE Corporation, 2020. URL: https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf.

[8] P. E. Kaloroumakis, M. J. Smith, Toward a Knowledge Graph of Cybersecurity Countermeasures, The MITRE Corporation, 2021. URL: https://d3fend.mitre.org/resources/D3FEND.pdf.

[9] I. Subach, O. Vlasenko, Architecture of intelligent SIEM for detecting cyber incidents in databases of military information and communication systems, Communication,

Informatization and Cybersecurity Systems and Technologies (4) (2023) 82–92. doi:10.58254/viti.4.2023.07.82 (in Ukrainian).

[10] H. I. Rybachok, Model of cyber incident neutralisation by military command authorities under conditions of limited time and response resources, Modern Information Technologies in the Sphere of Security and Defence 56 (2026) 166–176. URL: https://sit.nuou.org.ua/article/view/362350 (in Ukrainian).

[11] H. I. Rybachok, A methodology for assessing the effectiveness of cyber incident neutralization by military command authorities of the Armed Forces of Ukraine, Trudy Universytetu 4 (200) (2026) 66–72, in press (in Ukrainian).

[12] H. I. Rybachok, S. A. Mykus, Decision factors and decision context for military command authorities during cyber incidents, Cybersecurity: Education, Science, Technique (2026), accepted for publication (in Ukrainian).

Cross-Day Validation of FPR-Constrained Operating Thresholds for XGBoost-Based Botnet Detection

Vladyslav Samoilenko[1] and Sergii Gakhov[1]

[1] *Department of Cybersecurity Systems and Technologies, State University of Information and Communication Technologies, 7 Solomianska Str., Kyiv 03110, Ukraine*

Abstract

An FPR constraint is useful only if its calibrated threshold remains reliable beyond the benign traffic used for calibration. We evaluate cross-day robustness of frozen thresholds for a fixed Balanced 40-tree XGBoost botnet detector. The model is retrained on an early Friday window of CSE-CIC-IDS2018. Thresholds are calibrated on benign traffic from Thursday-15-02-2018 for FPR budgets of 0.25%, 0.5%, and 1%, then applied to three additional benign days and a later Friday holdout. The 0.25% and 0.5% constraints shifted the OOF optimum; at 1%, the unconstrained optimum already met the budget and the constraint was inactive. All empirical FPR values remained within budget; maximum other-day FPRs were 0.0861%, 0.2371%, and 0.5902%. The 0.5% budget gave the strongest compromise, with F1=0.99749 and 246.75 false alerts per 100,000 benign S2 flows. The results support cross-day portability within this benchmark, not a universal guarantee for arbitrary enterprise networks.

Keywords

botnet detection, XGBoost, false-positive rate, threshold calibration, cross-day robustness

1. Introduction

Machine-learning intrusion detectors can rank suspicious flows accurately, but operational value depends on false-alert load. At enterprise volumes, even a small benign alert rate can cause alert fatigue [1]. A threshold selected on one sample may fail as normal traffic changes, so it should be assessed through explicit FPR budgets on untouched benign traffic, not only F1 or area-under-curve metrics [2, 3].

Generalizability remains a weakness because benchmark results often reflect a narrow traffic generator or interval [2]. Time-aware validation reduces within-sequence leakage [4], while concept drift can alter legitimate activity after deployment [5]. For botnet detectors, benign tail scores determine false-alert load; the key question is whether a threshold calibrated on one benign day preserves its FPR budget on other days.

Our earlier time-ordered comparison identified XGBoost as the strongest candidate [9]. A subsequent study selected a 40-tree Balanced configuration using forward-validation quality, end-to-end throughput, and a 1% FPR constraint on one Thursday benign pool [10], leaving multi-day stability as future work. The present study does not repeat model selection, Pareto optimization, SMOTE analysis, or throughput benchmarking. Its novelty is a cross-day protocol: the configuration is fixed, three budgets are calibrated on one benign day, and each threshold is frozen before evaluation on several day-specific benign pools. The contribution is the protocol and measured budget compliance, irrespective of the direction of the external FPR change.

Corresponding authors

v.samoylenko@duikt.edu.ua (V. Samoilenko); gakhovsa@gmail.com (S. Gakhov)

0009-0001-6367-1755 (V. Samoilenko); 0000-0001-9011-8210 (S. Gakhov)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Experimental Protocol

Experiments use the CSE-CIC-IDS2018 TrafficForML/CICFlowMeter files [6]. The model is trained on the early Friday-02-03-2018 segment S1 (10:12-10:49 after a one-minute edge purge), containing 76,957 flows: 59,537 benign and 17,420 Bot. The later S2 segment (11:01-11:33) contains 123,792 flows: 56,332 benign and 67,460 Bot, and is used only as the temporally future holdout relative to S1. The binary label is one only for the exact Bot class. The fixed feature vector contains protocol, duration, total packets, total bytes, packets/s, bytes/s, and bytes/packet.

The fixed Balanced configuration uses 40 rounds, depth 3, learning rate 0.21577, subsample 0.75712, column subsample 0.79621, and histogram trees. The S1 class-weight ratio is 3.41774. Retraining uses seed 42 and one thread in the preceding study's Windows 10, Python 3.10.0, and XGBoost 3.1.3 environment. All results come from this run; differences from [10] reflect refitting and recalibration. Boosted trees suit structured flow data while retaining compact inference [7, 8].

Threshold calibration uses 996,077 benign-labelled flows from Thursday-15-02-2018. Cross-day evaluation uses only rows whose label is exactly Benign from Friday-16-02 (446,772 flows), Thursday-22-02 (1,048,213), and Thursday-01-03 (238,037). Attacks present in those files are excluded because FPR is defined over the negative class. Friday S2 benign flows are analysed separately: unlike the February and March-01 pools, S2 is temporally future relative to S1 but belongs to the same day as training. Thus, results on the additional days are described as cross-day robustness, not future temporal transferability.

For a benign pool d with N_d flows and model score p_i, the empirical FPR at threshold t is

$$\mathrm{FPR}_d(t)=\frac{1}{N_d}\sum_{i=1}^{N_d}\mathrm{I}\left(\hat{p}_i>t\right). \tag{1}$$

For each target budget α in {0.0025, 0.005, 0.01}, the minimum admissible threshold t_{min} is the (1-α) benign-score quantile on the calibration day, with a small numerical offset consistent with the strict decision rule $p_i>t$. Six chronological out-of-fold blocks from S1 are then used to select

$$t^*(\alpha)=\underset{t\geq t_{\min}(\alpha)}{\operatorname{argmax}}\ \mathrm{F1}_{\mathrm{OOF}}(t). \tag{2}$$

Each t*(α) is frozen and applied without adaptation to every validation pool. We report FPR, false alerts per 100,000 benign flows, the budget-compliance ratio $\mathrm{BCR}_d=\mathrm{FPR}_d/\alpha$, and the cross-day ratio $\mathrm{CDR}_d=\mathrm{FPR}_d/\mathrm{FPR}_{\mathrm{Thu}}$. A BCR not exceeding one denotes empirical budget compliance; CDR isolates a day-to-day change relative to the achieved calibration FPR. Wilson 95% intervals are reported descriptively because network flows are temporally dependent and therefore do not form independent Bernoulli trials.

3. Results and Discussion

Table 1 reports distinct thresholds for the three budgets. Constraints at 0.25% and 0.5% shifted the unconstrained OOF optimum; at 1%, that optimum (0.251221) already met the budget. No threshold equalled t_{min}, so constrained F1 maximization remained relevant.

Table 1
Frozen-threshold calibration and validation results (FPR and 95% CIs in percent)

α	t*	Thu-15 FPR [95% CI]	Max other-day FPR [95% CI]	S2 FPR	S2 P/R/F1
0.25	0.9916	0.2203 [0.2112–0.2297]	0.0861 [0.0807–0.0919]	0.098	0.9991/0.9318/0.9643
0.50	0.5539	0.4941 [0.4806–0.5081]	0.2371 [0.2279–0.2466]	0.247	0.9979/0.9971/0.9975
1.00	0.2512	0.9001 [0.8818–0.9189]	0.5902 [0.5758–0.6051]	0.609	0.9949/0.9982/0.9966

All empirical FPR values respected their budgets, and Thursday-15 was the most demanding benign pool at every threshold. Across auxiliary days, CDR ranged from 0.018–0.391 for α=0.25%, 0.031–0.480 for α=0.5%, and 0.203–0.656 for α=1%. Thus, benign upper-tail mass changed substantially across days, but downward relative to calibration. The maximum always occurred on Thursday-22 and remained below 60% of budget; even its upper Wilson bounds were compliant.

Expressed as budget utilization, the calibration-day BCRs were 0.881, 0.988, and 0.900 for α=0.25%, 0.5%, and 1%, respectively. The corresponding maximum auxiliary-day BCRs were 0.344, 0.474, and 0.590, whereas S2 used 0.390, 0.494, and 0.609 of the available budgets. These two ratios answer different questions: BCR indicates whether an operational limit is respected, while CDR measures movement relative to the achieved calibration FPR. Reporting both avoids interpreting a low external FPR as evidence of distributional stability when the calibrated threshold was already conservative. Here, compliance was accompanied by clear tail variation, so the result should be read as preservation of the selected operating limits rather than invariance of benign traffic.

The 0.5% budget offered the strongest compromise: on S2 it achieved precision 0.99794, recall 0.99705, and F1=0.99749, with 139 false positives and 199 false negatives (246.75 false alerts per 100,000 benign flows). Tightening to 0.25% cut false positives to 55 but raised false negatives to 4,600. Relaxing to 1% recovered 78 Bot flows relative to 0.5%, but raised false positives to 343 and slightly reduced F1 to 0.99657. Thus 0.5% is the best empirical balance here, although deployment utility may favor the higher-recall 1% setting.

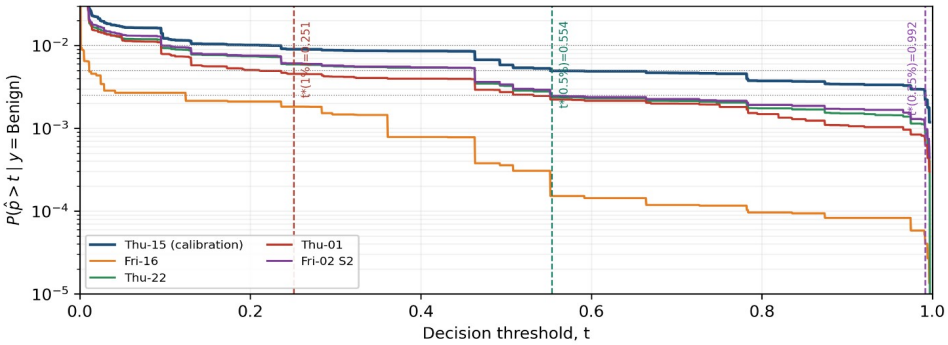


Figure 1: Cross-day upper-tail survival functions of benign XGBoost scores. The y-axis is logarithmic; vertical markers show the three frozen operating thresholds.

Figure 1 shows the largest benign upper-tail probability on Thursday-15 and markedly lower high-score probability on Friday-16; all other pools remained below calibration. At α=0.5%, the calibration estimate was 0.4941% and its Wilson upper bound 0.5081%; the method constrains empirical FPR, not the interval bound. The 1% row is therefore a compliance check, while the 0.25% and 0.5% rows demonstrate constraint-induced threshold movement.

Thursday-15 is a DoS GoldenEye/Slowloris day, although only rows labelled Benign were evaluated. Its heavier benign tail may therefore reflect scenario-specific background conditions rather than intrinsic conservatism of the method. This does not invalidate the frozen-threshold test, but it argues against calibrating on the quietest day, which could underestimate FPR. Future calibration should combine representative periods or include the most adverse observed benign tail, followed by untouched-day validation.

Freezing each threshold before examining the auxiliary pools is essential to the interpretation. If thresholds were recalibrated separately for every day, compliance would be built into the procedure and would not test portability of the operating point. The frozen design instead exposes changes in the benign score tail while holding the model and decision rule constant. In deployment, the protocol could be extended through periodic validation on adjudicated benign samples, monitoring of upper score quantiles, and predefined recalibration triggers. Such monitoring is necessary because the absence of violations in this benchmark does not imply that a fixed threshold will remain valid indefinitely.

4. Conclusions

This study extends an XGBoost operating-point method from one FPR budget and calibration day to three budgets and multiple frozen-threshold pools. All empirical budgets were preserved, while calibration consistently produced the largest false-alert rate. The 0.5% budget was most balanced, combining F1=0.99749 and recall=0.99705 on future S2 with an alert rate of 0.2468%. The results support cross-day threshold robustness within CSE-CIC-IDS2018 and show that multi-day validation reveals benign-tail variation even without budget violations.

All days share the same benchmark infrastructure and benign B-Profile mechanism, so the experiment does not establish cross-network portability. Wilson intervals are descriptive because flows are serially dependent, and the auxiliary days precede the Friday training window, so they are not future temporal transfer. Future work should calibrate on multiple benign days, optimize worst-case FPR, and evaluate genuine cross-network and prospective drift.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI ChatGPT for code generation, grammar and spelling checking, text rephrasing, and document formatting. After using this tool, the authors independently verified the experimental code and numerical results, reviewed and edited the content as needed, and take full responsibility for the publication's content.

References

[1] S. Tariq et al., Alert fatigue in security operations centres: Research challenges and opportunities, ACM Computing Surveys 57 (2025) 224. doi:10.1145/3723158.

[2] Z. A. Sheikh et al., Generalizability assessment of learning-based intrusion detection systems for IoT security: Perspectives of data diversity, Security and Privacy 8 (2025) e70014. doi:10.1002/spy2.70014.

[3] T. M. Silva Filho et al., Classifier calibration: A survey on how to assess and improve predicted class probabilities, Machine Learning 112 (2023) 3211–3260. doi:10.1007/s10994-023-06336-7.
[4] S. Liu, D. J. Zhou, Using cross-validation methods to select time series models: Promises and pitfalls, British Journal of Mathematical and Statistical Psychology 77 (2024) 337–355. doi:10.1111/bmsp.12330.
[5] M. A. Shyaa et al., Evolving cybersecurity frontiers: A comprehensive survey on concept drift and feature dynamics aware machine and deep learning in intrusion detection systems, Engineering Applications of Artificial Intelligence 137 (2024) 109143. doi:10.1016/j.engappai.2024.109143.
[6] Canadian Institute for Cybersecurity, CSE-CIC-IDS2018 dataset, University of New Brunswick, 2026. URL: https://www.unb.ca/cic/datasets/ids-2018.html (accessed 30 August 2026).
[7] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 785–794. doi:10.1145/2939672.2939785.
[8] M. R. A. Khan et al., Lightweight quantized XGBoost for botnet detection in resource-constrained IoT networks, IoT 6 (2025) 70. doi:10.3390/iot6040070.
[9] S. Gakhov, V. Samoilenko, Time-aware comparative analysis of machine learning methods for botnet activity detection in corporate networks, under review, 2026.
[10] V. Samoilenko, S. Gakhov, Method for high-throughput and reliable botnet detection via multi-objective optimization of XGBoost, under review, 2026.

Acoustic Anomaly Detection for Indoor Intrusion Using Ray-Traced Room Impulse Responses

Oleksandr Terletskyy[1], Valeriy Trushevskyy[1], and Ostap Hrytsyshyn[1]

[1] *Ivan Franko National University of Lviv, Universytetska 1, 79000 Lviv, Ukraine*

Abstract

Indoor intrusion detection based on sound is limited by two obstacles: real intrusion audio is scarce and expensive to record, and the sound reaching a microphone is strongly shaped by room geometry, surface materials, and source/receiver placement, so a model trained in one room or position generalizes poorly to another. This paper proposes an unsupervised acoustic anomaly-detection pipeline for indoor intrusion that reuses the authors' Unity-based ray-tracing sound-propagation simulator to (a) model the baseline acoustic signature — a frequency-resolved room impulse response (RIR) — of a protected room, (b) generate synthetic training data by convolving intrusion and background sounds with simulated RIRs across many source positions and frequency bands, and (c) score anomalies as deviations in per-band energy and reverberation features measured at the microphone position. Because only normal audio is required for training, the method follows the unsupervised anomalous-sound-detection paradigm established by the DCASE 2020 Challenge Task 2. The feature set (per-band energy, reverberation time, log-mel spectrogram, MFCC), a convolutional-autoencoder detector with one-class SVM and Gaussian mixture model baselines, and an evaluation plan based on AUC/pAUC metrics with simulation-to-reality validation are described. The proposed approach connects physics-based acoustic simulation with machine-learning-based intrusion detection, requiring only a single microphone and no labeled intrusion recordings.

Keywords

acoustic anomaly detection, indoor intrusion detection, room impulse response, ray tracing, synthetic training data, unsupervised anomalous sound detection

1. Introduction

Physical intrusion into homes, offices, and secure facilities is typically detected with cameras, passive-infrared (PIR) motion sensors, and contact or glass-break sensors. Cameras suffer from visual blind spots and darkness and raise privacy concerns; PIR and contact devices are line-of-sight or single-point sensors and can be bypassed. Acoustic monitoring is attractive because a single microphone can cover a whole room, and because break-in actions — glass breaking, door forcing, footsteps, movement of furniture — produce characteristic sounds. Commercial glass-break sensors already exploit the two-stage low-frequency "thud" followed by a high-frequency "shatter" signature.

However, learning-based acoustic intrusion detection faces two fundamental problems. First, real intrusion audio is rare and costly to collect, so labeled training data is scarce; staging break-ins is expensive and does not cover the diversity of real events. Second, the signal at the microphone is heavily colored by the room's geometry, surface materials, and the source and receiver positions. This causes severe domain shift between rooms and even between microphone placements within the same room, so a detector trained on recordings from one environment degrades in another.

Corresponding authors

oleksandr.terletskyi@lnu.edu.ua (O. Terletskyy); valeriy.trushevsky@lnu.edu.ua (V. Trushevskyy); ostap.hrytsyshyn@lnu.edu.ua (O. Hrytsyshyn)

0009-0001-7489-4428 (O. Terletskyy); 0000-0001-8672-7152 (V. Trushevskyy); 0009-0000-4875-2330 (O. Hrytsyshyn)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

This paper addresses both problems by exploiting a physics-based acoustic simulator that the authors have built and validated in prior work [1, 2, 3]. The simulator, implemented in the Unity engine, traces sound rays through a 3D reconstruction of the protected room, splits sound into frequency bands, and produces frequency-resolved room impulse responses in real time. It is proposed to use this simulator as (i) the generator of the room's baseline acoustic fingerprint and (ii) the generator of positionally and spectrally diverse synthetic training data, and to cast intrusion detection as unsupervised anomaly scoring on deviations from the simulated baseline.

2. Related work and scientific novelty

Prior work on acoustic intrusion detection divides into two families. Passive event detection classifies sounds such as glass breaking, footsteps, and door slams using mel-frequency cepstral coefficients (MFCC) or log-mel spectrogram features with convolutional neural networks (CNN), Gaussian mixture models (GMM), or one-class support vector machines (SVM) [10]. Active acoustic sensing emits a probe signal and analyzes echoes, Doppler shift, or changes in the room transfer function. A representative active system, AudioGuard [9], captures the periodic Doppler shift caused by an intruder's walking motion using only a commodity speaker–microphone pair and reports miss rates of 0% (line-of-sight) and 1.75% (non-line-of-sight) with a 4.17% false alarm rate.

In machine listening, unsupervised anomalous sound detection (ASD) — training only on normal sound and flagging deviations — was established as a benchmark by the DCASE 2020 Challenge Task 2 [4], built on the MIMII [5] and ToyADMOS datasets and evaluated with the area under the ROC curve (AUC) and partial AUC (pAUC) over a low false-positive-rate range. Follow-up datasets such as MIMII DUE [6] explicitly target domain shift, which is the central difficulty for room-dependent intrusion audio. Autoencoder-based detectors trained on normal data remain the standard baseline for this task [11]. Separately, synthetic and simulated room impulse responses (RIRs) generated by geometric acoustics — image-source and ray-tracing methods — are widely used to augment scarce, position-dependent training data for sound event localization and detection [7, 8] and for scene synthesis in domestic sound event detection [12].

The novelty of the present work is to unite these threads. A ray-tracing propagation simulator is used not merely for generic augmentation, but as the generator of the protected room's baseline acoustic fingerprint and of positionally and frequency-band-resolved synthetic intrusion data, and intrusion detection is cast as unsupervised anomaly scoring on per-band energy and reverberation deviations. This directly extends the authors' ray-tracing simulator, which features frequency banding via Butterworth bandpass filters, Maekawa diffraction modeling with binary-search optimization, frequency-based path caching, and Sabine-validated reverberation [1, 2], as well as the subsequent real-time static-ray impulse-response method explicitly aimed at audio-based anomaly detection [3].

3. Proposed method

3.1. Baseline acoustic fingerprint

The protected room is reconstructed as a 3D scene in Unity with per-surface material absorption coefficients. The authors' engine traces 20–100 rays per frame with up to 8 bounces, splits sound into 9–10 frequency bands using Butterworth bandpass filters, applies Maekawa diffraction with binary-search optimization and frequency-based path caching, and produces a frequency-resolved impulse response h_b(t) per band b at the microphone position, at a real-time cost of approximately 4–16 ms per frame on a Ryzen 7 7800X3D [1, 3]. The set of band impulse responses, together with

its reverberation descriptors — per-band reverberation time T60, early decay time (EDT), and band-energy ratios — forms the empty-room baseline fingerprint.

The physical fidelity of the fingerprint is supported by the prior validation of the simulator against Sabine's formula:

$$T_{60} = \frac{0.161 \cdot V}{A} \quad (1)$$

where V is the room volume and A is the total absorption; simulated reverberation times approached the Sabine prediction as the number of rays increased [1, 3].

3.2. Synthetic training-data generation

A library of normal indoor sounds (HVAC, speech, appliances, ambient street noise) and candidate intrusion sounds (glass break, door forcing, footsteps, drilling) is convolved with simulated RIRs at many source positions — windows, doors, and points along plausible walking paths — and at the fixed microphone position:

$$y_b(t) = x(t) * h_b(t) \quad (2)$$

where x(t) is a dry source signal and h_b(t) is the simulated band-b impulse response for a given source position. Varying source position, occupancy, absorption coefficients, and additive noise yields a large, positionally and spectrally diverse dataset without any field recording, mitigating both data scarcity and room/position domain shift.

3.3. Feature extraction

For each analysis window, the following features are computed:

- per-band energy over the 9–10 ray-tracing frequency bands, matching the simulator's banding;
- reverberation features: per-band T60, EDT, and direct-to-reverberant energy ratio;
- standard log-mel spectrogram (16 kHz, 128 mel bins) and MFCC features for compatibility with DCASE-style models [4].

3.4. Anomaly model and scoring

Only normal features are used to train the detector, following the unsupervised ASD paradigm [4, 11]. A convolutional autoencoder is the primary model; one-class SVM and GMM serve as lightweight baselines. The anomaly score is the reconstruction error (respectively, the negative log-likelihood); a threshold set on held-out normal data flags intrusion events. Because the fingerprint is band- and position-resolved, deviations can be attributed to particular frequency bands and approximate source regions, providing interpretable alerts (e.g., a high-frequency transient near a window is consistent with glass breakage).

3.5. Active-sensing option

When a loudspeaker is available, a known chirp probe can be emitted periodically and the measured RIR compared with the simulated baseline. A change in the room transfer function — caused by a new absorbing or scattering body such as a person — raises the anomaly score even for a silent intruder, complementing the passive pipeline in the spirit of active systems such as AudioGuard [9].

4. Evaluation plan and expected results

It is expected that (i) models trained purely on ray-traced synthetic data will detect held-out intrusion events (glass break, door forcing, footsteps) convolved with unseen RIRs, and (ii) synthetic RIR augmentation will improve robustness to microphone and room domain shift relative to a raw-audio baseline, in line with results reported for simulated-RIR augmentation in sound event localization and detection [7, 8].

Evaluation will use AUC and pAUC — the latter computed over a low false-positive-rate range [0, p] with p = 0.1, following DCASE 2020 Task 2 [4] — together with event-based precision/recall and false alarms per hour, which is the operationally relevant metric for security systems. Simulation-to-reality transfer will be validated on a small set of real recordings made in a room whose geometry is modeled in the simulator; the previously demonstrated agreement of simulated reverberation times with Sabine's prediction supports the physical fidelity of the fingerprint [1, 3]; the accuracy–cost trade-off observed in that validation is summarized in Table 1. Domain-shift robustness will additionally be probed by perturbing absorption coefficients and microphone position between training and test conditions, analogously to the domain-shift protocol of MIMII DUE [6].

Table 1
Reverberation time T60 estimated by the ray-tracing simulator for a 10 m × 20 m × 20 m room versus the number of traced rays, compared with the Sabine prediction of 5.37 s [1, 3]

Number of rays	Ray-traced T60 (s)	Error vs. Sabine 5.37 s (%)
10	2.1	61
100	3.1	42
200	3.5	35
500	3.9	27
1000	4.1	24

5. Conclusions

This paper has specified an unsupervised acoustic anomaly-detection pipeline for indoor intrusion that turns the authors' ray-tracing sound-propagation simulator into a generator of room acoustic fingerprints and positionally diverse synthetic training data, addressing the twin obstacles of scarce intrusion audio and room/position domain shift. The main result of this contribution is a concrete, physically grounded architecture — frequency-band RIR fingerprinting, RIR-convolution data synthesis, and reconstruction-error anomaly scoring — that requires only normal-room audio for training and connects the authors' acoustic-simulation line of work [1, 2, 3] to practical cyber-physical facility security. Future work is the empirical simulation-to-reality evaluation of the pipeline and its deployment on resource-constrained edge devices.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to: translate certain text fragments into English, perform grammar and spelling checks, and paraphrase or

reword content. After using these tools, the authors carefully reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] O. Terletskyy, V. Trushevskyy, Ray tracing based approach for the sound propagation simulation, Visnyk of the Lviv University, Series Applied Mathematics and Informatics 32 (2024) 73–85.

[2] O. Terletskyy, V. Trushevskyy, Rozrobka prohramnoho zabezpechennia dlia symuliatsii akustychnykh khvyl za dopomohoiu trasuvannia promeniv [Software development for acoustic wave simulation using ray tracing], in: ITSec: Information Technology Security, Proceedings of the 13th International Scientific-Technical Conference, Ivan Franko National University of Lviv, Lviv, 2024, pp. 196–197.

[3] O. Terletskyy, V. Trushevskyy, Real-time acoustic impulse response generation using static ray tracing for anomaly detection, in: SMICS-2025: Proceedings of the International Scientific and Technical Conference, Ivan Franko National University of Lviv, Lviv, 2025, pp. 161–166. URL: https://smics2025.lnu.edu.ua/proceedings/.

[4] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, N. Harada, Description and discussion on DCASE2020 Challenge Task2: Unsupervised anomalous sound detection for machine condition monitoring, in: Proceedings of the 5th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), 2020, pp. 81–85. arXiv:2006.05822.

[5] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, Y. Kawaguchi, MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection, in: Proceedings of the 4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), 2019, pp. 209–213.

[6] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, Y. Kawaguchi, MIMII DUE: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions, in: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2021, pp. 21–25.

[7] C. Ick, B. McFee, Leveraging geometrical acoustic simulations of spatial room impulse responses for improved sound event detection and localization, 2023. arXiv:2309.03337.

[8] Y. Koyama, K. Shigemi, M. Takahashi, K. Shimada, N. Takahashi, E. Tsunoo, S. Takahashi, Y. Mitsufuji, Spatial data augmentation with simulated room impulse responses for sound event localization and detection, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8872–8876.

[9] T. Wang, Z. Li, H. Yan, et al., AudioGuard: Omnidirectional indoor intrusion detection using audio device, ACM Transactions on Internet of Things 5 (2024) 1–26.

[10] Z. Mnasri, S. Rovetta, F. Masulli, Anomalous sound event detection: A survey of machine learning based methods and applications, Multimedia Tools and Applications 81 (2022) 5537–5586.

[11] R. Müller, F. Ritz, S. Illium, C. Linnhoff-Popien, Acoustic anomaly detection for machine sounds based on image transfer learning, 2020. arXiv:2006.03429.

[12] N. Turpault, R. Serizel, A. Parag Shah, J. Salamon, Sound event detection in domestic environments with weakly labeled data and soundscape synthesis, in: Proceedings of the 4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), 2019, pp. 253–257.

Методика виявлення фішингових електронних повідомлень на основі гібридних алгоритмів машинного навчання

Ігор Аверічев[1], Петро Поночовний[1]

[1] *Державний університет інформаційно-комунікаційних технологій, солом'янська 7, 03110, Київ, Україна*

Анотація

У тезі розглянуто проблему виявлення фішингових електронних повідомлень як одного з основних векторів кібератак на корпоративні та хмарні інфраструктури. Проаналізовано існуючі методи захисту (превентивні, технічні, на основі машинного навчання та реагування на інциденти). Запропоновано методику детекції на основі гібридної моделі (наївний баєсівський класифікатор + SVM) з уніфікованим інтерфейсом інтеграції. Наведено математичний апарат оцінювання (теорема Баєса, метрики Accuracy, Precision, Recall, F1-score). Показано доцільність комплексного підходу Defense-in-Depth для підвищення ефективності кіберзахисту систем зв'язку та хмарної інфраструктури.

Ключові слова

фішинг, електронна пошта, машинне навчання, гібридна модель, Naive Bayes, SVM, детекція загроз, кіберзахист, Defense-in-Depth, F1-score, URL-аналіз, управління інцидентами.

1. Актуальність

Проблеми захисту інформаційних систем від фішингових атак зумовлена постійним зростанням кількості кіберінцидентів, спрямованих на компрометацію облікових даних користувачів через електронну пошту. Фішингові повідомлення залишаються одним із найпоширеніших векторів початкового проникнення в корпоративні та хмарні інфраструктури. Традиційні методи фільтрації на основі сигнатур і чорних списків демонструють обмежену ефективність проти нових і модифікованих атак. У зв'язку з цим актуальним є застосування алгоритмів машинного навчання для побудови адаптивних систем детекції [1; 2; 3].

2. Мета роботи

Полягає у розробленні методики виявлення фішингових електронних повідомлень із використанням гібридного підходу на основі алгоритмів машинного навчання та уніфікованого інтерфейсу інтеграції, що забезпечує підвищення точності класифікації та зручність впровадження в існуючі системи кіберзахисту.

Типовий процес фішингової атаки включає кілька послідовних етапів: планування (вибір цілі та збір адрес), налаштування інфраструктури (створення підроблених сайтів і шаблонів листів), безпосередню атаку (масова розсилка), збір скомпрометованих даних та фінальну експлуатацію отриманої інформації (рис. 1).

Corresponding authors
iaverichev19@gmail.com (I. Averichev); petja9186@gmail.com (P. Ponochovnyi)
0009-0008-9766-0115 (I. Averichev); 0009-0008-6480-6990 (P. Ponochovnyi)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

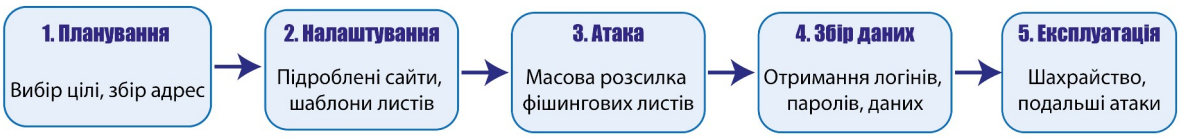


Рис. 1: Етапи фішингової атаки

Аналіз методів захисту від фішингових атак дозволяє виділити чотири основні групи: превентивні, технічні (фільтраційні), методи на основі машинного навчання та методи реагування на інциденти [1; 5]. Превентивні заходи включають навчання користувачів, політики безпеки організації, протоколи автентифікації відправника (SPF, DKIM, DMARC) та білі списки. Технічні фільтри базуються на чорних списках URL/IP-адрес, контентному аналізі, перевірці вкладень антивірусними сканерами та репутаційних сервісах. Методи машинного навчання використовують класифікатори (Naive Bayes, SVM, Random Forest, нейронні мережі) для аналізу текстових і структурних ознак повідомлень [4]. Методи реагування передбачають ізоляцію підозрілих листів, сповіщення SOC, блокування шкідливих доменів та проведення форензики інциденту.

Ефективний кіберзахист будується за принципом глибокого захисту (Defense-in-Depth), що поєднує превенцію, детекцію, реагування та відновлення. Класифікацію методів і ланцюг захисту схематично представлено на рисунку 2.



Рис. 2: Класифікація методів захисту від фішингу та ланцюг кіберзахисту

Для реалізації запропонованої методики застосовано гібридну модель, що поєднує наївний баєсівський класифікатор та метод опорних векторів (SVM) [3; 4]. Такий підхід дозволяє ефективно обробляти текстові ознаки повідомлень та структурні характеристики URL-адрес.

Ймовірність того, що повідомлення є фішинговим за умови наявності певного слова (ознаки), обчислюється за теоремою Баєса (1):

$$P(phishing|word)=\frac{P(phishing)\cdot P(word|phishing)}{P(phishing)\cdot P(word|phishing)+P(legit)\cdot P(word|legit)}, \quad (1)$$

де:

$P(\text{phishing} \mid \text{word})$ — апостеріорна ймовірність того, що лист є фішинговим за наявності ознаки «word»;

$P(\text{phishing})$ — апріорна ймовірність фішингового повідомлення;

$P(\text{word} \mid \text{phishing})$ — ймовірність появи ознаки у фішингових листах;

$P(\text{legit})$ — апріорна ймовірність легітимного повідомлення;

$P(\text{word} \mid \text{legit})$ — ймовірність появи ознаки у легітимних листах.

Архітектура рішення включає етапи попередньої обробки тексту (видалення стоп-слів, токенізація, лематизація), вилучення ознак (характеристики доменів, наявність IP-адрес у URL, розбіжності між текстом посилання та атрибутом href, ключові слова соціальної інженерії) та класифікацію [2; 4]. Для забезпечення інтеграції з різними клієнтськими застосунками реалізовано уніфікований програмний інтерфейс.

Ефективність моделей оцінювалася за стандартними метриками класифікації (2; 3; 4; 5):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

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

$$Recall = \frac{TP}{TP + FN}, \tag{4}$$

$$Recall = 2 \cdot \frac{Precision \cdot Recall}{Precisio + Recall}, \tag{5}$$

де:

$TP\left(True\ Positive\right)$ — правильно виявлені фішингові повідомлення;

$TN\left(True\ Negative\right)$ — правильно класифіковані легітимні повідомлення;

$FP\left(False\ Positive\right)$ — хибні спрацювання (легітимний лист позначено як фішинг);

$FN\left(False\ Negative\right)$ — пропущені фішингові повідомлення.

Експериментальне оцінювання проводилося на наборі даних із легітимними та фішинговими URL-адресами. Порівнювалися моделі Multinomial Naive Bayes, Logistic Regression, Random Forest, Decision Tree, KNN, SVC та Multilayer Perceptron. Гібридний підхід продемонстрував найбільш збалансовані значення метрик. Особливу увагу приділено зниженню частки хибнонегативних спрацювань.

Практичне впровадження методики передбачає використання сервісу на базі високопродуктивного бекенду з асинхронною обробкою запитів, що забезпечує прийнятний час реакції системи навіть за умов значного навантаження.

3. Висновки

Розроблена методика виявлення фішингових електронних повідомлень на основі гібридних алгоритмів машинного навчання підвищує ефективність детекції порівняно з окремими класичними моделями. Дослідження методів захисту підтвердило необхідність комплексного підходу (превенція — детекція — реагування). Уніфікований інтерфейс інтеграції спрощує впровадження рішення в існуючі системи захисту інформації та

управління інцидентами. Отримані результати можуть бути використані для подальшого вдосконалення засобів кіберзахисту систем зв'язку та хмарної інфраструктури.

Список використаних джерел:

[1] Alshamrani A., Myneni S., Chowdhary A., Huang D. A survey on phishing attacks and countermeasures for email security. IEEE Access. 2024. Vol. 12. P. 112345–112370.

[2] Sahingoz O. K., Buber E., Demir O., Diri B. Machine learning based phishing detection from URLs. Expert Systems with Applications. 2023. Vol. 215. Article 119302.

[3] Kumar S., Jain A., Bansal A. Machine learning approaches for phishing email detection: A comparative study. Applied Soft Computing. 2024. Vol. 151. Article 111095.

[4] Chen L., Zhao Y., Sun J. Hybrid phishing detection using NLP and ensemble learning techniques. Information Sciences. 2024. Vol. 646. P. 119–134.

[5] Аверічев, І., Роженко, А., & Кихтенко, Є. (2025). ІННОВАЦІЙНІ ПІДХОДИ ДО ПІДВИЩЕННЯ РІВНЯ КІБЕРБЕЗПЕКИ КОРПОРАТИВНИХ МЕРЕЖ ПРИ ВИКОРИСТАННІ ХМАРНИХ ТЕХНОЛОГІЙ. Електронне фахове наукове видання «Кібербезпека: освіта, наука, техніка», 1(29), 732–747. https://doi.org/10.28925/2663-4023.2025.29.934



TRACK 2. METHODS OF COUNTERING DISINFORMATION INFLUENCE AND SOCIAL ENGINEERING

Eduard B. Fleißner and his textbook “Handbuch der Kryptographie”

Oleg Gutik[1], Mykola Shcherbyna[1] and Yurii Boberskyy[1]

[1] *Ivan Franko National University of Lviv, Universytetska 1, 79000 Lviv, Ukraine*

Abstract

We give a short review of the biography of the Austrian military cryptographer Eduard Freiherr von FLEIßNER von WOSTROWITZ, his textbook "Handbuch der Kryptographie", and his cipher FLEIßNER's Grille. Also, we discuss on contribution of Eduard Freiherr von FLEIßNER von WOSTROWITZ in the development of the cryptology.

Keywords

Cryptography, History of Cryptography, Cryptanalysis, Military Cryptography

1. Early Life and Military Career

Eduard Freiherr (Baron) von Fleißner von Wostrowitz, the son of an Austrian cavalry captain, was born on January 25, 1825, in Lviv (then Lemberg), which was then part of the Austrian Empire. After training as an officer at the renowned Theresian Military Academy in Wiener Neustadt (50 km south of Vienna) — founded in 1752 by Empress Maria Theresa and still operating today as an academy for Austrian officers — he became a second lieutenant in the Imperial and Royal Austrian Army in 1843. He continued to rise through the ranks, subsequently being appointed commander of the school for brigade commanders in 1870 and acting commander of an allied division in 1872. He retired in 1876 and moved to Vienna in 1880, where he died on April 29, 1888.

While serving as the commander of various military units, and especially as an instructor at the school for brigade commanders, he encountered cryptographic tools and methods. He later completed a book on cryptography, which he felt was urgently needed. It was published in Vienna in 1881 under the title Handbuch der Kryptographie (Handbook of Cryptography) [2].

2. Fleißner’s Contributions to Cryptography

What contribution did Fleißner make to modern cryptography? His main contribution was the work itself, which found tangible expression in a book that contained all the experience he had gathered during his service. If we look at the seminal works on cryptography from the last decade of the 19th century and the first decade of the 20th century ([1, 4, 8]), there are few comprehensive manuals quite like Fleißner’s.

Handbuch der Kryptographie contains much that was valuable for its time. Its most striking quality is the clear and systematic way in which the author organized the material and individual systems. Furthermore, he demonstrated a strong understanding of the difference between encryption, decryption, and unauthorized decryption — what are known today as cryptography and cryptanalysis, respectively. The author announced a second volume of his book dedicated to

Corresponding authors

oleg.gutik@lnu.edu.ua (O. Gutik); mykola.shcherbyna@lnu.edu.ua (M. Shcherbyna); yurii.boberskyy@lnu.edu.ua (Y. Boberskyy)

0000-0001-8513-0282 (O. Gutik); 0009-0003-9761-9466 (M. Shcherbyna); 0009-0004-0073-2217 (Y. Boberskyy)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

cryptanalysis, but it was never published. It was here that Fleißner reached his limits — not in knowledge, but regarding the political and military environment of the era. The time was not yet ripe for public, scientific cryptanalysis.

Looking at Fleißner's book, which is nearly 150 years old, it is surprising how relevant his preface remains when translated into modern colloquial language. For example, in the first paragraph, he writes:

> "Thanks to the introduction of postcards and the fact that encrypted telegrams are permitted in the cryptography of private correspondence or the art of encryption and decryption, a science that until then had existed only for a few classes, has gained significance and interest for the wider public as well"

One need only replace the phrases "postcards" and "encrypted telegrams" with "emails" and "data transmission," respectively, and this paragraph would perfectly suit a book titled On the Cryptography of the Past Century.

The same applies to the third chapter, where Fleißner states that general knowledge of cryptography is insufficient, despite its frequent use by government agencies and professions that should be familiar with the subject. As a result, they often use encryption methods that are ineffective at protecting secrets.

3. The Structure of His Work and the Stencil Cipher

Fleißner divided his book into the following three parts:

1. General Issues in Cryptography and Preferred Methods
2. The New Lattice/Transposition Cipher (Patronen-Geheimschrift)
3. The Art of Unauthorized Decryption (Cryptanalysis)

The first and third parts provide a detailed overview of the tools and methods in the field, based on the knowledge of the time. From today's perspective, the second part — which proposes a new transposition cipher called Patronen-Geheimschrift (stencil cipher) — is of particular interest.

The term Patrone (stencil) refers to a square piece of cardboard with holes cut into it. These holes are arranged so that when the cardboard is rotated 90 degrees around its center into one of four possible positions, the holes never overlap. The four sides, each fifteen fields long, were labeled 1 through 4 (or A through D), and on the reverse side, 5 through 8 (or E through H). The nature of the cryptogram in the central hole indicates which surface and side to use when starting the encryption and decryption processes.

In the introduction to Part II, Fleißner lists nine advantages of his new cipher. In the fifth point, he claims:

> "It has an extraordinary level of security, superior to other methods of encryption. Science and art cannot find the key, except by a stroke of luck, which is possible for any cipher."

He also emphasized the enormous variety of possible keys. However, as we know today, Fleißner was mistaken twice regarding his new cipher:

- First, the variety of keys depends heavily on the length of the square's sides and is further limited by the strict mathematical constraints necessary to allow the square to rotate around its center. Further restrictions arise when considering equivalent and weak keys.
- Second, the key and the cipher device are nearly identical. Therefore, the Patrone itself must be kept secret — a condition that severely undermines the modern cryptographic axiom that the cipher device/algorithm should not need to be kept secret (Kerckhoffs's principle).

Despite these flaws, Fleißner modestly stated in his foreword that he wrote the book for non-professionals and general readers, rather than for experts.

4. The Structure of His Work and the Stencil Cipher

To wrap up this topic, it is worth asking what we can learn from Fleißner's work. There are three main takeaways to consider:

1. Do not assume that cryptographic needs and requirements are known, understood, and accepted everywhere.
   The remarks in Fleißner's preface illustrate this quite clearly.
2. Do not overestimate the security of your own system.
   Fleißner's Patronen-Geheimschrift serves as an excellent negative example of this overconfidence.
3. Carefully consider what should be kept secret from a security perspective, and what should not.

Two examples illustrate this. First, Fleißner's Patrone should not have had to be a secret, since it acted as the encryption device, not just the key. Second, weak encryption methods do not become more secure simply by attempting to keep their inner workings secret.

If we keep these principles in mind, Fleißner's work was not in vain, and his contribution to modern cryptography continues to bear fruit.

5. Historical Context and World War I Usage

The use of a rotating lattice points back to the late 18th century, as these devices were quite popular at the time. The most notable early work in this regard is K. F. Hindenburg's 1796 paper [5], titled "Fragen eines Ungenannten über die Art, mit dem Gitter geheim zu schreiben" (Questions from an Anonymous Author on the Method of Writing Secretly with a Grid). This is the oldest complete description of the rotating grid, written almost a century before Fleißner's well-known version [3].

Ultimately, Fleißner's Patronen-Geheimschrift was frequently used by the Austrian army during World War I. The use of this cipher on the Italian front is described in a dedicated monograph [2], noting that the Fleißner grid (Fleißner Grille) used was sized 15 × 15. However, despite Fleißner's earlier claims of unbreakability, this cipher was successfully broken by the Italian engineer Luigi Sacco [6, 7].

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 3.1 Pro in order to: Grammar and spelling check. After using this tool, authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] F. L. Bauer, Decrypted secrets: methods and maxims of cryptology, 2nd ed., Springer, 2000.
[2] C. Colavito and F. Cappellano, The secret war on the Italian front in WWI (1915 - 1918), 3rd English ed., Roma, 2021.
[3] Eduard B. Fleißner von Wostrowitz, Handbuch der Kryptographie: Anleitung zum Chiffriren und Dechiffriren von Geheimschriften, L. W. Seidel, 1881.
[4] J. S. Galland, An historical and analytical bibliography of the literature of cryptology, Northwestern University Press, 1945.
[5] C. F. Hindenburg, Fragen eines Ungenannten Ungenannten über die Art Dutch Gitter geheim zu schreiben,Archiv der reinen und angewandten Mathematik. Ill (1795), 347-351.
[6] L. Sacco, Manuale di crittografia, Roma, 1926.
[7] L. Sacco, Manuale di crittografia - terza edizione aggiornata e aumentata, Roma, 1947.
[8] D. Shulman, An annotated bibliography of cryptography, New York, Garland Publishing, Inc., 1976.

Predictive Cyber Threat Intelligence: Combining NLP-Based Entity Extraction with Graph Link Prediction

Yevheniia Ivanchenko[1], Iryna Lozova[1] and Polina Halushka[1]

[1] *State University of Information and Communication Technologies, Solomianska 7, 03110 Kyiv, Ukraine*

Abstract

Cyber threat intelligence is published as a large volume of unstructured textual reports, which makes manual correlation of threat actors, malware, vulnerabilities, and affected products difficult to scale. This paper presents a pipeline that combines a locally deployed large language model for automated entity extraction with classical graph algorithms for predictive analysis. Eleven real, publicly documented CTI reports describing three well-known vulnerabilities (Log4Shell, EternalBlue, and ProxyLogon) were processed to build a heterogeneous knowledge graph of actors, malware, vulnerabilities, products, and MITRE ATT&CK techniques. Degree and betweenness centrality were used to identify the most critical nodes, while the Adamic-Adar index was applied to predict undocumented actor-vulnerability relationships. A leave-one-out validation procedure achieved a hit-rate@3 of 75%, demonstrating that the proposed approach can recover real, previously removed relationships from the remaining graph structure alone.

Keywords

cyber threat intelligence, knowledge graph, named entity recognition, large language models, link prediction, Adamic-Adar index, MITRE ATT&CK

1. Problem Statement

Modern organizations are exposed to a continuous stream of cyber threat intelligence (CTI) published as free-form text: vendor security advisories, national CERT bulletins, and research blog posts. Each individual report typically describes only a fragment of the overall threat landscape – a single vulnerability, a single malware sample, or a single observed campaign. A complete picture of "who is attacking whom, with what tools, and through which vulnerabilities" only emerges after an analyst manually reads and cross-references dozens of such fragments.

This manual correlation process does not scale. As the volume of published CTI grows, security analysts spend an increasing share of their time on repetitive extraction and cross-referencing tasks rather than on decision-making. Moreover, purely manual analysis is reactive by nature: it documents relationships between actors and vulnerabilities only after they have already been publicly reported, leaving little room for anticipating which not-yet-documented relationships are likely to emerge next.

The general problem addressed in this work is therefore twofold: (1) how to automatically transform unstructured CTI text into a structured representation suitable for computational analysis, and (2) how to use that structured representation not only to describe already known relationships, but also to produce and validate predictions about relationships that have not yet been observed.

2. Scientific Novelty

Existing approaches to AI-assisted cyber threat analysis largely fall into two categories. The first category uses large language models (LLMs) as end-to-end risk classifiers: a single artifact (a file

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

hash, an IP address, or a short incident description) is passed to an LLM, which returns a qualitative risk assessment. While useful for triage automation, this approach evaluates artifacts in isolation and provides no mechanism for reasoning about relationships between multiple artifacts, actors, or vulnerabilities. The second category relies on static, manually curated knowledge bases, such as the MITRE ATT&CK matrix [8], which describe known adversary behavior but are not automatically populated from raw text and do not offer a quantitative predictive capability.

The approach proposed in this paper differs from both categories in three respects. First, entity extraction is performed directly on unstructured natural-language CTI reports by a locally deployed, freely available LLM, removing the dependency on pre-structured threat feeds or paid cloud APIs. Second, the architecture explicitly separates two responsibilities that are often conflated in LLM-centric prototypes: the LLM is responsible only for fact extraction (NLP), while all analytical reasoning – identifying critical nodes and predicting new relationships – is delegated to well-established, deterministic graph algorithms (degree/betweenness centrality and the Adamic-Adar index [1]). Third, and most importantly, the quality of the resulting predictions is evaluated quantitatively rather than described only qualitatively: a leave-one-out validation protocol measures what fraction of real, previously observed actor-vulnerability relationships the model could have predicted from the remaining graph alone. This type of ground-truth-based evaluation is typically absent from LLM-based CTI triage prototypes, which usually report only a subjective severity score with no accuracy metric attached to it.

3. Proposed Solution

The proposed solution is implemented as a five-stage Python pipeline that transforms unstructured CTI reports into a validated set of predictive insights.

3.1. Data and Entity Model

The evaluation corpus consists of eleven short CTI reports describing three real, publicly documented vulnerabilities: CVE-2021-44228 [6] (Log4Shell, affecting Apache Log4j), CVE-2017-0144 [7] (EternalBlue, affecting Windows SMBv1), and CVE-2021-26855 [9] (ProxyLogon, affecting Microsoft Exchange Server). The report texts are based on publicly available attribution and analysis from CISA [3], Microsoft MSTIC [4], and CrowdStrike [5], and reference real, documented threat actors, including HAFNIUM, Sandworm (APT44), Lazarus Group, and AQUATIC PANDA. Each report is modeled as a source of four entity types: threat actor, malware, CVE identifier, and affected product.

3.2. Pipeline Architecture

For each report, a locally hosted LLM (Ollama, llama3.2:3b) is prompted to return a structured JSON object containing the four entity types described above. To improve robustness against the imprecise output that small local models occasionally produce (for example, string placeholders such as "null" instead of a true empty value), the extracted entities are passed through a sanitization step before being added to the graph. The extracted entities from all eleven reports are then merged into a single directed knowledge graph using the NetworkX library [2], with typed nodes (actor, malware, cve, product, technique) and typed edges (uses, exploits, affects, implements). Edges of type implements connect malware nodes to MITRE ATT&CK techniques through a static lookup table, separating fact extraction from the enrichment step.

3.3. Graph Analysis and Link Prediction

Two families of graph metrics are computed on the resulting knowledge graph. Degree and betweenness centrality are used to rank nodes by structural importance, identifying which vulnerabilities act as the most critical points in the threat landscape – that is, which nodes are connected to the largest number of distinct actors or serve as a structural "bridge" between otherwise separate clusters of attack activity. Second, for every actor-CVE pair that is not yet directly connected, the Adamic-Adar index is computed as a similarity score based on shared neighbors (shared malware or shared attack techniques); pairs with the highest score are treated as the most likely candidates for a future, not-yet-documented relationship.

Figure 1 shows the resulting knowledge graph obtained from the eleven input reports: 31 nodes and 30 edges, colored by entity type (actors in red, malware in orange, CVE identifiers in purple, affected products in blue, and MITRE ATT&CK techniques in green).

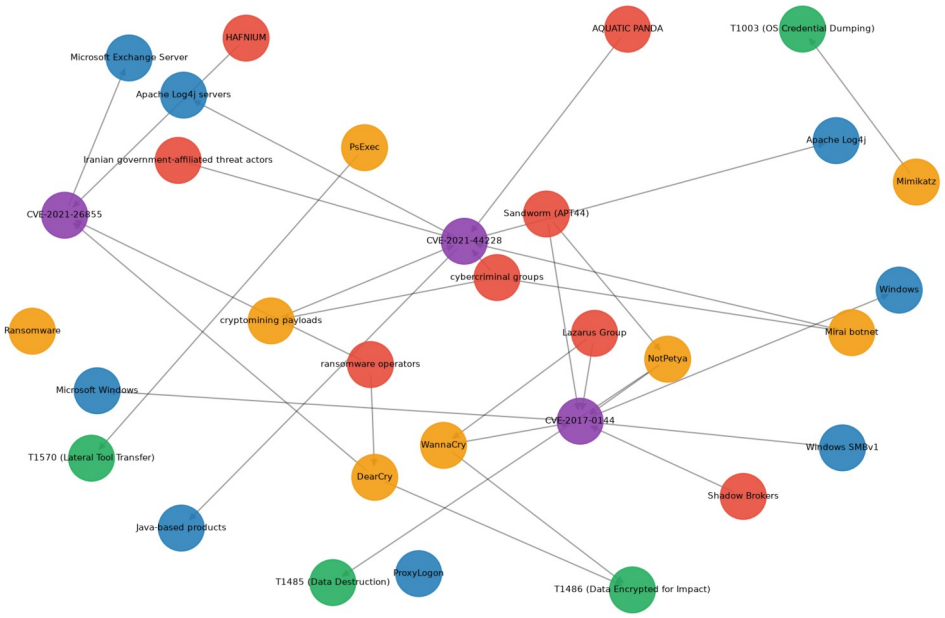


Figure 1: Knowledge graph constructed from eleven CTI reports describing Log4Shell, EternalBlue, and ProxyLogon.

3.4. Validation Methodology

The predictive quality of the link-prediction step is assessed using a leave-one-out procedure: for every real, observed actor-CVE relationship, that edge is temporarily removed from the graph, the Adamic-Adar ranking is recomputed on the remaining structure, and the outcome is recorded as a hit if the removed relationship reappears among the top-3 predicted candidates. This produces a quantitative hit-rate@k metric rather than a subjective assessment of prediction quality, and was found to be sensitive to the density of the underlying graph: an earlier, sparser version of the corpus produced a hit-rate@3 inflated by tied zero scores, which was corrected by enriching the corpus with an additional report describing a genuinely shared tool between two actors.

4. Conclusions

This paper presented a pipeline for predictive cyber threat intelligence that combines automated, LLM-based entity extraction from unstructured text with classical, deterministic graph algorithms for structural analysis and link prediction. Applied to eleven real, publicly documented reports covering three well-known vulnerabilities (Log4Shell, EternalBlue, and ProxyLogon), the pipeline produced a 31-node, 30-edge knowledge graph in which CVE-2017-0144 (EternalBlue) was identified as the most structurally critical vulnerability, connecting two independent attack campaigns (WannaCry, attributed to Lazarus Group, and NotPetya, attributed to Sandworm/APT44).

The Adamic-Adar-based link-prediction step, validated through a leave-one-out protocol, achieved a hit-rate@3 of 75%, meaning that six out of eight real actor-CVE relationships could have been correctly anticipated from the remaining graph structure alone, before they were explicitly documented in the input reports. This confirms that separating fact extraction (delegated to an LLM) from analytical reasoning (delegated to graph algorithms) is a viable architecture for predictive CTI analysis, and that its predictive quality can and should be measured quantitatively rather than assumed.

The main limitation of the current implementation is the size of the evaluation corpus, which is sufficient to demonstrate the method but too small to yield a graph with rich, non-trivial neighbor structure for every candidate pair. Future work includes evaluating the pipeline on a substantially larger CTI corpus, replacing the static leave-one-out evaluation with a temporal, chronologically ordered validation scheme, and comparing the small local model used in this work with larger cloud-hosted LLMs in terms of extraction accuracy and downstream prediction quality.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) to assist in debugging the Python implementation of the described pipeline (graph construction, leave-one-out validation procedure). After using these tools/services, the author(s) reviewed and edited the content as necessary and bear full responsibility for the content of this publication.

References

[1] L. A. Adamic, E. Adar, Friends and neighbors on the web, Social Networks 25 (2003) 211–230. doi:10.1016/S0378-8733(03)00009-1.

[2] A. Hagberg, D. Schult, P. Swart, Exploring network structure, dynamics, and function using NetworkX, in: Proceedings of the 7th Python in Science Conference (SciPy2008), 2008, pp. 11–15.

[3] CISA, Iranian government-sponsored APT actors compromising federal networks, exfiltrating data, Cybersecurity Advisory AA22-320A, 2022. URL: https://www.cisa.gov/news-events/cybersecurity-advisories/aa22-320a.

[4] Microsoft Threat Intelligence Center (MSTIC), HAFNIUM targeting Exchange servers with 0-day exploits, Microsoft Security Blog, 2021. URL: https://www.microsoft.com/en-us/security/blog/2021/03/02/hafnium-targeting-exchange-servers/.

[5] CrowdStrike Intelligence, AQUATIC PANDA conducts espionage operation targeting academic institutions, CrowdStrike Blog, 2022. URL: https://www.crowdstrike.com/blog/aquatic-panda-conducts-espionage-operation/.

[6] National Vulnerability Database, CVE-2021-44228 Detail, NIST, 2021. URL: https://nvd.nist.gov/ vuln/detail/CVE-2021-44228.
[7] National Vulnerability Database, CVE-2017-0144 Detail, NIST, 2017. URL: https://nvd.nist.gov/ vuln/detail/CVE-2017-0144.
[8] MITRE, MITRE ATT&CK, 2023. URL: https://attack.mitre.org.
[9] National Vulnerability Database, CVE-2021-26855 Detail, NIST, 2021. URL: https://nvd.nist.gov/ vuln/detail/CVE-2021-26855.

Human Competencies as a Risk Mitigation Factor in Disinformation and Social Engineering Attacks

Anna Kiełbus[1]

[1] *University of the National Education Commission, Podchorążych 2, 30-084 Kraków, Poland*

Abstract

The increasing use of artificial intelligence in disinformation and social engineering attacks creates new challenges for organisational cybersecurity. Phishing, vishing, impersonation, deepfakes and AI-generated deceptive content can exploit authority, urgency, trust and insufficient information verification. This paper proposes the 4K Human Risk Model (4K-HR Model), integrating four 21st-century competencies—critical thinking, creativity, communication and cooperation—with cybersecurity risk assessment. The model extends conventional assessment based on probability and impact by introducing human vulnerability as an additional risk dimension. Five representative scenarios are used to demonstrate the approach. The paper provides a conceptual basis for further empirical research on the relationship between competencies, human vulnerability and cyber resilience.

Keywords

cybersecurity, disinformation, social engineering, human factor, risk management, critical thinking, 21st-century competencies, phishing, deepfake, artificial intelligence

1. Introduction

Cybersecurity threats increasingly exploit human behaviour as well as technical vulnerabilities. Social engineering manipulates individuals into revealing information, granting access or performing an action that benefits an attacker. ENISA reports that phishing, including vishing, malspam and malvertising, was the leading initial intrusion vector in its 2025 assessment and notes the growing use of large language models to improve phishing and automate social engineering [1].

Generative AI further changes the attack environment. Synthetic text, voice and video can make fraudulent communications more convincing, while deepfake impersonation can exploit trusted organisational relationships [2], [3]. Recent research therefore calls for human-centred cybersecurity approaches combining behavioural, organisational and technical dimensions [4]. This paper asks how 21st-century competencies can be incorporated into the assessment and mitigation of human-related cyber risk.

2. Human Factor and 21st-Century Competencies

Lamri identifies four competencies important in the contemporary professional environment: creativity, communication, critical thinking and cooperation [5]. In cybersecurity, critical thinking supports credibility assessment, source verification and questioning of authoritative instructions. Creativity facilitates alternative explanations and response options. Communication enables reporting and independent verification, while cooperation supports collective decision-making and reduces isolated responses.



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

These competencies are treated as potential risk modifiers rather than direct determinants of secure behaviour. Vulnerability also depends on context, including time pressure, stress, organisational culture, procedures, perceived authority and attack characteristics.

3. Disinformation and Social Engineering as Human-Centred Risks

Disinformation and social engineering differ in purpose but may exploit similar cognitive and behavioural mechanisms. The OECD Truth Quest Survey shows that distinguishing false or misleading online content can be difficult and is influenced by contextual and presentation factors [6]. A systematic review of employee awareness of social engineering identified perceptual, organisational, social, psychological, behavioural, cultural, security-awareness and content-related factors [7].

AI-supported attacks reduce the reliability of traditional credibility cues. Attackers can combine impersonation, urgency and confidentiality with AI-generated text, synthetic voice or deepfake video. Effective mitigation therefore requires technical controls together with procedures that strengthen verification, communication and cooperation [3].

4. Risk Management and Human Vulnerability

Information-security risk is commonly represented as $R = P \times I$, where P is probability and I is impact. ISO/IEC 27005:2022 provides guidance for risk identification, assessment and treatment [8], while NIST CSF 2.0 supports organisational processes for understanding and prioritising cybersecurity risk [9].

For social engineering, successful exploitation also depends on human vulnerability. The proposed extension is:

$$R_H = P \times I \times V$$

where:

P – probability of the attack,

I – potential impact,

V – human vulnerability.

V represents vulnerability in context. It is not a permanent characteristic of an employee and may be influenced by competencies, time pressure, organisational procedures and attack characteristics.

5. The 4K Human Risk Model

The proposed 4K-HR Model links an attack and its manipulation technique with human vulnerability, decision-making, consequences and mitigation. The four competencies form the human layer of the model.

The model assumes that stronger 4K competencies may reduce vulnerability; this relationship remains a research proposition requiring empirical validation. The model complements, rather than replaces, technical and organisational controls.

6. Human Vulnerability Index

A preliminary Human Vulnerability Index (HVI) is proposed. Each competency is assessed on a five-point scale from 1 (very low vulnerability) to 5 (very high vulnerability). The dimensions are CT—critical thinking, CR—creativity, CO—communication and CP—cooperation.

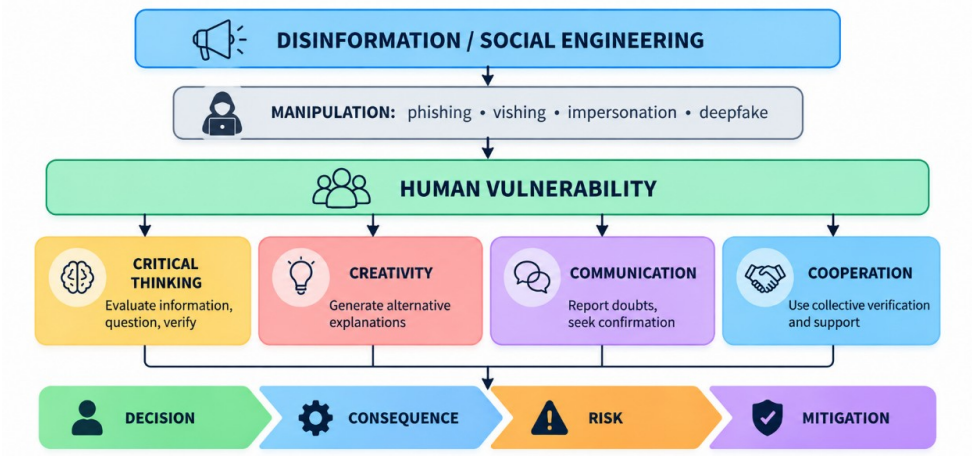


Fig. 1. Conceptual structure of the 4K Human Risk Model (4K-HR Model).

Using equal weights in the initial conceptual model:

$$HVI = (CT + CR + CO + CP) / 4$$

A higher HVI indicates greater vulnerability. Equal weighting is provisional because empirical evidence is insufficient to justify different weights. Future studies should calibrate the index using expert assessment, behavioural experiments or statistical modelling.

7. Scenario-Based Risk Assessment

The following values are illustrative rather than empirically validated. P, I and V use five-point scales; the final column identifies the competencies most directly relevant to the scenario. Organisational disinformation is discussed qualitatively below.

Scenario	P / I / V → R	4K focus
Phishing	4 / 3 / 3 → 36	CT
Spear-phishing	4 / 4 / 4 → 64	CT + CO
Vishing	3 / 4 / 4 → 48	CO + CP
AI impersonation / deepfake	3 / 5 / 4 → 60	CT + CP
Organisational disinformation	4 / 4 / 3 → 48	CT + CR

Fig. 2. Illustrative application of the 4K-HR model.

8. AI-Based Impersonation: Example and Mitigation

An employee receives an apparently authentic request from a senior manager: “I am currently in a meeting. Please urgently execute the transfer specified below. The transaction is confidential.” The attacker combines authority, urgency, impersonation and AI-generated content.

The model prompts four checks: question the request (critical thinking); consider alternative explanations such as compromise or impersonation (creativity); verify through another channel (communication); and consult an authorised colleague (cooperation). Practical controls may include independent verification, a two-person principle for high-risk transactions, incident reporting and targeted exercises. Research on differential phishing susceptibility also supports considering individual and contextual differences [10].

9. Discussion, Limitations and Future Research

The model treats the employee as an active security resource and connects 21st-century competencies with risk management. It shifts the focus from general security awareness towards security capability: knowing that phishing exists does not guarantee correct behaviour during a sophisticated attack. For organisations, the proposed logic is Risk → Required competency → Expected behaviour → Control.

The model is conceptual and has not been empirically validated. The causal relationship between 4K competencies and resistance to social engineering remains to be established, and HVI weights are provisional. Future research should combine 4K assessment with controlled phishing simulations, deepfake scenarios, behavioural measurements and statistical analysis to test whether HVI predicts cybersecurity behaviour and whether competency development improves cyber resilience.

10. Conclusions

Disinformation and social engineering show that cybersecurity cannot rely exclusively on technological controls. The 4K-HR Model integrates critical thinking, creativity, communication and cooperation with risk assessment and introduces human vulnerability as an additional dimension. The model proposes these competencies as potential risk-mitigation factors while explicitly recognising the need for empirical validation. It provides a basis for further research on human-centred cyber resilience and for designing risk-based competency development programmes.

References

[1] European Union Agency for Cybersecurity (ENISA), ENISA Threat Landscape 2025, Heraklion, Greece: ENISA, 2025.

[2] R. Jabir, J. Le, and C. Nguyen, “Phishing Attacks in the Age of Generative Artificial Intelligence: A Systematic Review of Human Factors,” AI, vol. 6, no. 8, art. 174, 2025, doi: 10.3390/ai6080174.

[3] K. T. Pedersen, L. Pepke, T. Stærmose, M. Papaioannou, G. Choudhary, and N. Dragoni, “Deepfake-Driven Social Engineering: Threats, Detection Techniques, and Defensive Strategies in Corporate Environments,” Journal of Cybersecurity and Privacy, vol. 5, no. 2, art. 18, 2025, doi: 10.3390/jcp5020018.

[4] K. Khadka and A. B. Ullah, “Human factors in cybersecurity: an interdisciplinary review and framework proposal,” International Journal of Information Security, vol. 24, art. 119, 2025, doi: 10.1007/s10207-025-01032-0.

[5] J. Lamri, Kompetencje XXI wieku: kreatywność, komunikacja, krytyczne myślenie, kooperacja, Warszawa, Poland: Wolters Kluwer, 2021.

[6] OECD, The OECD Truth Quest Survey: Methodology and Findings, OECD Digital Economy Papers, no. 369, Paris, France: OECD Publishing, 2024, doi: 10.1787/92a94c0f-en.

[7] M. Alomair, T. Issa, S. Zaung Nau, and B. Abu Salih, “The key factors that influence employees’ awareness of social engineering: A systematic literature review,” Heliyon, vol. 11, no. 16, e44012, 2025, doi: 10.1016/j.heliyon.2025.e44012.

[8] International Organization for Standardization, ISO/IEC 27005:2022: Information security, cybersecurity and privacy protection—Guidance on managing information security risks, Geneva, Switzerland: ISO, 2022.

[9] C. Pascoe, S. Quinn, and K. Scarfone, The NIST Cybersecurity Framework (CSF) 2.0, NIST Cybersecurity White Paper CSWP 29, Gaithersburg, MD, USA: National Institute of Standards and Technology, 2024, doi: 10.6028/NIST.CSWP.29.

[10] M. Canham, S. Dawkins, and J. Jacobs, “Not All Victims Are Created Equal: Investigating Differential Phishing Susceptibility,” National Institute of Standards and Technology, 2024.

A mathematical model for assessing the cyber resilience of an information resource with regard to input data adequacy

Oleksandra Kovalchuk[1], Eugenia Ivanchenko[1]

[1] *State University of Information and Communication Technologies, Solomyanska 7, 03110 Kyiv, Ukraine*

Abstract

This paper proposes a mathematical model for the dynamic assessment of the cyber resilience of an information resource, addressing the challenge of comprehensive quantitative security measurement under real threat conditions and incomplete input data. The model is based on a tuple representation of the protected object, formalizes interdependencies between assets through a dynamic graph, and introduces a mechanism for logistic damping of metrics that accounts for data adequacy. The integral Cyber Resilience Index (CRI) is computed through weighted hierarchical aggregation of indicators across the six functions of the NIST CSF 2.0, with the possibility of adaptive weight-coefficient updating, and is decomposed into partial estimates of confidentiality, integrity, availability, and data adequacy.

Keywords

cyber resilience, information resource, data adequacy, CRI index, NIST CSF 2.0, cascading degradation, dynamic model

1. Introduction

The rapid digitalization of critical infrastructure and the growing complexity of multi-stage cyber threats, systematized in the MITRE ATT&CK Framework [1], require a shift from reactive protection to proactive cyber-resilience assessment. Traditional approaches based on static compliance audits (ISO/IEC 27001, NIST SP 800-53) [2, 3] capture the security posture only at a single moment and do not reflect risk evolution.

Existing methodologies have two main limitations: (1) a predominant focus on qualitative maturity assessment without a formalized quantitative measure of cyber resilience; (2) quantitative models treat input data as fully reliable, ignoring incompleteness, timeliness and trustworthiness [4, 5], which systematically distorts results. An additional gap is the insufficient modelling of cascading degradation among interdependent assets.

Consequently, there is a need for a mathematical model that combines the NIST CSF 2.0 process approach, quantitative evaluation of technical and organizational characteristics, explicit accounting for input-data adequacy, cascading asset degradation, and adaptive updating of risk weights.

The aim of this study is to develop such a model, treating data adequacy as a full-fledged calculation parameter. The objectives include formalizing the information resource, defining indicators aligned with NIST CSF 2.0 functions, assessing data adequacy, modelling cascading degradation, and constructing the integral CRI index with its decomposition to support managerial decisions.

Corresponding authors

oa.kovalchuk@duikt.edu.ua (O. Kovalchuk); e.ivanchenko@duikt.edu.ua (E. Ivanchenko)

0009-0003-1218-7810 (O. Kovalchuk); 0000-0003-3017-5752 (E. Ivanchenko)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Mathematical Model for Cyber Resilience Assessment

The model is built as a four-level hierarchy corresponding to the taxonomy of NIST CSF 2.0 [6] (function → category → subcategory → metric) and rests on three key ideas: a tuple representation of the protected object, explicit consideration of data adequacy, and modeling of the temporal dynamics of asset states.

2.1. Formalization of the Protected Object

The information resource is represented by the tuple:

$$M_{IR} = \langle A, G(Qdata|i(t)), V, T, S, P, D \rangle, \quad (1)$$

where A is the set of assets (hardware, software, information, personnel), V and T are vulnerabilities and threats, S denotes protective measures, $G(Q_{\text{data}})$ is the dynamic interdependency graph whose edge weights are modulated by data adequacy, and P, D are processes and regulatory documentation. The state of each asset evolves according to

$$x_i(t|1) = x_i(t) - d_i(t)x_i(t) + r_i(t)[1 - x_i(t)] + \sum_j q_i(t)x_j(t), \quad (2)$$

where $x_i(t) \in [0|1]$ is the resilience level of asset i, $d_i(t)$, $r_i(t)$ are the degradation and recovery intensities, and $a_{ij}(t) = a^0_{ij} \cdot Q_{data,j}(t)$, is the influence of asset j on i modulated by the local data-adequacy of the source asset. A projection operator $x_i(t+1) \leftarrow \min\{1, \max\{0, x_i(t+1)\}\}$ enforces the bounds after each update.

2.2. Input data adequacy and logistic damping of metrics

For each metric source, an integral adequacy indicator $Q_i(t) \in [0, 1]$ is computed as a weighted linear convolution of six data quality parameters — relevance, completeness, timeliness, accuracy, sufficiency, and authenticity:

$$Q_{data}(t) = \sum_k \alpha_k \cdot q_k(t), \sum \alpha_k = 1, \quad (3)$$

where $Q_i(t)$ — integral data-adequacy indicator of the i-th metric source at time t; $q_k(t) \in [0|1]$ — k-th partial quality parameter (relevance, completeness, timeliness, accuracy, sufficiency, authenticity); α_k — weight of the k-th quality parameter. A distinction is made between the local indicator $Q_{data,i}(t)$, used for damping a specific metric, and the aggregated indicator $Q_{global}(t) = \sum_i \omega_i Qdata,_i(t)$, which reflects the overall trust in the information environment.

To avoid distortion from unreliable sources the corrected metric is obtained by logistic damping:

$$m_i^*(t) = m_i \cdot \frac{1}{1 + e^{-k(Q_{data}(t) - \tau)}} \quad (4)$$

where m_i is the nominal metric value; τ=0.4 is the unreliability threshold; k is the slope steepness of the logistic curve; $Q_{data}(t)$ — local data-adequacy indicator of the metric source. The base parameters τ = 0.4 and k = 15 were derived analytically from a significance level α = 0.05 and a transition-zone width Δ = 0.2. When $Q_i \to 1$ the metric is fully trusted; when $Q_i < \tau$ it is isolated ($m_i^* \to 0$). The same damping is applied to graph edges, implementing cascading isolation of distorted information.

2.3. Cascading degradation in the dynamic dependency graph

Security breaches and reduced data adequacy propagate bottom-up through the graph $G(Q_{\text{data}})$ (from hardware to personnel). Edge weights are adjusted by the current adequacy of the source-asset data:

$$w_{ij}^{*}(t) = w_{ij} \cdot \frac{1}{1+e^{-k(Q_{data,i}(t)-\tau)}} \quad (5)$$

where w_{ij}^{*} is the dynamic (adequacy-modulated) weight of the edge from asset j to asset i; w_{ij} is the nominal structural dependency weight; $Q_{data,j}(t)$ is the local data-adequacy indicator of the source asset.

Under full data adequacy (Q_{data} = 1), the dependency weight equals the nominal value w_{ij}, whereas a critical drop in the quality of information about the state of the source asset ($Q_{data} \to 0$) damps the connection ($w_{ij}^{*} \to 0$). This mechanism isolates distorted information and signals higher management levels about a loss of transparency in the given segment of the information resource, preventing incorrect cascading propagation of unreliable estimates between interdependent assets. As a heuristic criterion for bounding the strength of cascading interactions, the condition $\rho(A(t)) < 1$ is used, where $\rho(A(t))$ is the spectral radius of the asset interdependency matrix.

2.4. Integral Cyber Resilience Index (CRI)

Cyber resilience is defined as the system's ability to preserve its target operability function P(t) under attack conditions. For a nominal trajectory $P_{nom}(t) \equiv 1$, the integral index takes the normalized form:

$$CRI = \frac{1}{T}\int_{t_0}^{t_1} P(t)\,dt, \quad T¿ t_1 - t_0\ , \quad (6)$$

where $CRI(t_0|t_1)$ – integral Cyber Resilience Index over the observation interval; $P(t)$ – instantaneous cyber-resilience indicator (operability function); $P_{nom}(t) \equiv 1$ – nominal (ideal) operability trajectory. The instantaneous cyber-resilience indicator P(t) is obtained by successive hierarchical aggregation of the corrected metrics:

$$S_s(t) = \sum_{i \in s} \omega_i m_i^{*}(t), F_k(t) = \sum_{c \in C_k} \beta_c C_c(t), \quad (7)$$

$$F_k(t) = \sum_{c \in C_k} \beta_c C_c(t), \quad (9)$$

$$C_c(t) = \frac{1}{|\mathrm{SubCa\,t_c}|} \sum_{s \in \mathrm{SubCa\,t_c}} S_s(t), P(t) = \sum_{k=1}^{6} W_k(t) F_k(t), \quad (8)$$

$$P(t) = \sum_{k=1}^{6} W_k(t) F_k(t), \quad (10)$$

where $S_s(t)$ is the score of subcategory s, $C_c(t)$ is the score of category c, $F_k(t)$ is the score of the k-th NIST CSF 2.0 function (Govern, Identify, Protect, Detect, Respond, Recover), $m_i^{*}(t)$ are the adequacy-corrected metrics, ω_i and β_c are the local weights of metrics and categories respectively (obtained by AHP [7, 8] and normalized to unity), and $W_k(t)$ are the (possibly adaptive) weights of the six functions satisfying $\sum_k W_k(t) = 1$. The weights allow adaptive adjustment depending on the current risk profile R(t), formed from telemetry indicators:

(11)

where W_k^{0} is the base AHP weight, δ_k the sensitivity of function k, and $\Delta R(t)$ the aggregated deviation of the risk profile from its baseline. This allows the model to increase the weight of Respond or Recover during an active incident while preserving normalization.

2.5. Decomposition of the CRI Index and Operational Interpretation

To ensure interpretability of results, the index CRI(t) is decomposed into a vector of partial estimates $P(t) = (P_C(t), P_I(t), P_A(t), P_Q(t))$, corresponding to confidentiality, integrity, availability, and data adequacy. Each estimate is computed as a normalized weighted sum of corrected metrics, with $P_Q(t) = Q_{global}(t)$, which eliminates duplication between the partial adequacy estimate and the integral trust indicator. Joint interpretation of the pair (CRI, Q_{global}) allows the operational states shown in Table 1 to be distinguished.

Table 1
Operational interpretation of the CRI(t) index in the {CRI × Q_{global}} space

CRI(t)	$Q_{global}(t)$	Operational state	Characteristics
≥ 0.85	≥ 0.80	Managed resilience	System stable, data trustworthy
≥ 0.85	0.70–0.80	Resilience with limited trust	High resilience, insufficient data trust
0.50–0.85	≥ 0.70	Controlled degradation	Reduced security, monitoring still reliable
0.50–0.85	0.40–0.70	Degradation with limited visibility	Insufficient data to localize degradation
< 0.50	0.40–0.70	Uncontrolled degradation	Critical drop in resilience and data trust
any value	< 0.40	Undetermined	Assessment unreliable due to low data adequacy

This classification forms a complete partition of the phase space {CRI × Q_{global}}: every pair of values CRI(t) and $Q_{global}(t)$ corresponds to exactly one operational state, making it possible to distinguish formally high resilience obtained from unreliable data from a genuinely well-managed system state.

3. Conclusions

The proposed model provides a quantitative dynamic assessment of cyber resilience that explicitly accounts for input-data adequacy and cascading asset degradation. The integral CRI index and its decomposition into confidentiality, integrity, availability and data-adequacy dimensions enable timely detection of security degradation and support decision-making at operational, managerial and strategic levels. The model can be used for continuous compliance assessment, budget justification, comparative analysis of distributed resources and verification of incident-response plans.

References

[1] MITRE Corporation. MITRE ATT&CK Framework. URL: https://attack.mitre.org/
[2] ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection – Information security management systems – Requirements.
[3] NIST Special Publication 800-53 Rev. 5. Security and Privacy Controls for Information Systems and Organizations. NIST, 2020.
[4] Pendleton M., Garcia-Lebron R., Cho J.-H., Xu S. A Survey on Systems Security Metrics. ACM Computing Surveys. 2016. Vol. 49, No. 4. P. 1–35.

[5] Wang R. Y., Strong D. M. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. 1996. Vol. 12, No. 4. P. 5–33.
[6] The NIST Cybersecurity Framework (CSF) 2.0. NIST, 2024. URL: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf
[7] Saaty T. L. The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation. New York: McGraw-Hill, 1980. 287 p.
[8] Saaty T. L. Decision Making with the Analytic Hierarchy Process. International Journal of Services Sciences. 2008. Vol. 1, No. 1. P. 83–101.
[9] Poolsappasit N., Dewri R., Ray I. Dynamic Security Risk Management Using Bayesian Attack Graphs. IEEE Trans. Dependable Secure Comput. 2012. Vol. 9, No. 1. P. 61–74.
[10] Ross R. S., Graubart R., Bodeau D., McQuaid R. Systems Security Engineering: Cyber Resiliency Considerations. NIST SP 800-160 Vol. 2 Rev. 1. NIST, 2021.
[11] Cybersecurity Capability Maturity Model (C2M2). Version 2.1. U.S. Department of Energy, 2022. URL: https://www.energy.gov/ceser/cybersecurity-capability-maturity-model-c2m2
[12] National Cyber Security Centre (NCSC). Cyber Assessment Framework. URL: https://www.ncsc.gov.uk/collection/caf
[13] Federal Information Processing Standards Publication 199 (FIPS PUB 199). Standards for Security Categorization of Federal Information and Information Systems. NIST, 2004.

Velocity Features as the Structural Invariant of Card-Testing Attacks in Online Payment Systems

Yurii Kuzhii[1] and Yurii Furgala[1]

[1] *Ivan Franko National University of Lviv, Drahomanova 50, 79005 Lviv, Ukraine*

Abstract

Anti-fraud systems of online payment services are built on feature spaces assembled empirically — dozens to hundreds of attributes without a functional typology. We show that such spaces are structurally non-uniform and that a large part of their features carries no independent signal at all. The transaction feature space is organized into four functional families (temporal, velocity, amount-structure, identity) and their separability with respect to the fraud label is measured on 11,127 real card transactions by MANOVA, the silhouette index, sequential family ablation and stratified tests. The velocity family dominates (silhouette 0.81 against 0.06 for temporal features) and alone reproduces the quality of the full space (PR-AUC 0.551 against 0.547), while its removal costs 61 % of PR-AUC. Popular temporal features turn out to be indirect proxies of velocity behavior: the significant "weekend" effect (χ^2 11.73, $p < 0.001$) vanishes under stratification to the high-velocity subsample (χ^2 0.56, $p = 0.46$). The identified composite signature "at least two card transactions per hour with an amount repeated within a 10-minute window" holds an 80.6 % fraud rate at a 5.14 % base rate and covers 32.7 % of all fraud in 2.1 % of the volume, giving a compact detection rule for card-testing attacks. External validation on IEEE-CIS (590,540 transactions) confirms the structural principle but shows that the dominant behavioral family is regime-specific.

Keywords

card-testing attacks, payment fraud detection, behavioral patterns, feature engineering, social engineering

1. Introduction

Payment fraud in online services is a mass-scale attack surface rather than a set of isolated incidents. In 2023 Brazil alone recorded about 3.7 million fraudulent card transactions with aggregate losses of about 3.5 billion BRL, while the National Bank of Ukraine reports an annual growth of card-fraud losses. A substantial share of these losses begins with *card-testing*: an automated verification of stolen card credentials through a series of small transactions, which precedes the monetization of the compromised data and is documented in industry reports of Visa, Stripe and Adyen as one of the most widespread attacks against payment gateways. The credentials it verifies come predominantly from phishing and other social-engineering campaigns or from data breaches, so detecting the testing stage cuts the monetization path of those attacks.

Detection of such attacks is a pattern-recognition task in a high-dimensional feature space. In practice these spaces are built empirically: dozens to hundreds of attributes — time of day, day of week, amount, merchant, country, device — without a systematic typology. This overloads the model and invites interpretation errors, when a significant signal of one feature is in fact an indirect manifestation of a deeper behavioral variable. The industry belief about fraudster activity "at night" or "on weekends" is the typical case: such time slots merely coincide with periods of increased velocity activity of a compromised card.

Corresponding author

yurii.kuzhii@lnu.edu.ua (Yu. Kuzhii)

0009-0001-7890-8575 (Yu. Kuzhii); 0000-0001-7316-7424 (Yu. Furgala)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The problem addressed here is therefore not the accuracy of another classifier but the *structure* of that feature space: which family carries the class separation, and which families only reproduce it indirectly.

2. Novelty with Respect to Known Works

Classical work on card fraud treats feature engineering as auxiliary. Dal Pozzolo et al. [1] calibrate probabilities under severe class imbalance (fraud rate < 1 %); Jurgovsky et al. [2] apply LSTM to card transaction sequences; Lucas et al. [3] develop multi-perspective hidden Markov models that implicitly encode short-term patterns. Bahnsen et al. [4] give the first systematic study of feature-engineering strategies, aggregating attributes over time windows for a 1.5–3-fold quality gain; in APATE [5] card–merchant network features capture the graph structure of fraud propagation. On Brazilian data, Carneiro et al. [6] note repeated amounts as an artifact of stolen-card testing, while De Sá et al. [7] and Carcillo et al. [8] tune classifiers to the market's loss function and to streaming operation.

In all of these works features are *added*; which of them are structurally redundant is not asked. Card-testing itself, though well documented by payment systems, is rarely a separate analytical category in the literature: its markers (inter-transaction gaps, amount repetition) are detected among dozens of others, without emphasis on their invariance.

The novelty of this work is threefold. First, the transaction feature space is organized into functional families *a priori*, by the nature of the information a feature reflects, and separation strength is measured per family, not per feature. Second, the redundancy of popular temporal features is established not by importance ranking but by a stratified test that removes the velocity effect and shows the temporal signal to be a hidden common effect. Third, a single composite feature operationalizes the card-testing signature at an 80.6 % fraud rate — a rule usable alongside a model. The same family-wise methodology was previously applied by the authors outside the financial domain, to role-based patterns of Counter-Strike 2 players [9].

3. Feature Space and Methods

The public Predict Chargeback Frauds dataset [10] is used: 11,127 real transactions of 1–30 May 2015 with a masked card number, a second-precision timestamp, an amount and a chargeback label; 9,260 unique cards, fraud rate 5.14 % (N = 572). The transaction-to-card ratio of 1.20 indicates predominantly single-use cards, the many-transaction ones being tied to fraudulent activity.

Twenty-six features are grouped into four families by the nature of the information they reflect: **temporal** (7) — cyclically encoded hour and day of week, *weekend*, *night*, *afternoon_peak*; **velocity** (10) — the gap to the card's previous transaction, cumulative and rolling counts and sums in 1-, 6- and 24-hour windows, repetition of the previous amount and of an amount within a 10-minute window; **amount-structure** (7) — logarithm, roundness modulo 10/50/100, fractional part, thresholds; **identity** (2) — Bayesian-smoothed target encodings of the BIN and the amount. Velocity features respect causality: for the i-th transaction of a card they use only prior transactions of the same card, and identity encodings are fitted on the training part only.

The space was analyzed by PCA; univariate ANOVA and the Mann–Whitney U test; MANOVA (Wilks' λ) per family; the silhouette index [12] for geometric separation; logistic regression and gradient boosting on a time-ordered 75/25 split with sequential family ablation and bootstrap intervals for PR-AUC (1000 resamples); permutation importance; and a χ^2 stratification test for hidden common effects.

4. Results

Structural non-uniformity of the space. All four families are statistically significant by MANOVA ($p < 10^{-30}$), which at N = 11,127 is uninformative on its own, so the conclusions rest on effect size. By that criterion the families differ by an order of magnitude: velocity (Wilks' $\lambda = 0.682$) and identity (0.659) against temporal (0.986) and amount (0.975), where λ close to 1 means that within-group variation overrides the between-group one. The silhouette index gives the same picture in a sharper form: velocity +0.81, identity +0.65, amount +0.22, temporal +0.06 (Fig. 1). The index of the identity family is inflated, since target-encoded features carry direct information about the label; the honest measure of natural behavioral separability is therefore the velocity index. The space is not one-dimensional: the first two principal components explain only 37.2 % of the variance, so the behavior is not reducible to a single "fraud score".

Ablation. Sequential removal of families (Table 1) confirms the hierarchy. Removing velocity is catastrophic: PR-AUC falls from 0.547 to 0.212 (–61 %), while removing any other family costs at most 6 %. The velocity family alone reproduces the full space (PR-AUC 0.551 against 0.547, precision in the top 1 % of risk 0.926, recall in the top 5 % even higher, 0.511 against 0.454) — the remaining 16 features add no independent signal. Temporal features alone are close to a random classifier (PR-AUC 0.073 at a base rate of 0.062). Bootstrap 95 % intervals separate without-velocity 0.212 [0.165; 0.276] from full space 0.547 [0.478; 0.616] and velocity only 0.551 [0.474; 0.630], which almost fully overlap. Gradient boosting is qualitatively identical (0.422 full, 0.178 without velocity, 0.535 velocity only). Permutation importance is consistent: *bin_te* (identity) ranks first at ΔPR-AUC 0.256 but is inflated by target-encoding memorization, and the strongest purely behavioral features are *gap_min_prev* (velocity) 0.190 and *n_tx_card_1h* 0.039, against at most 0.006 for every temporal feature.

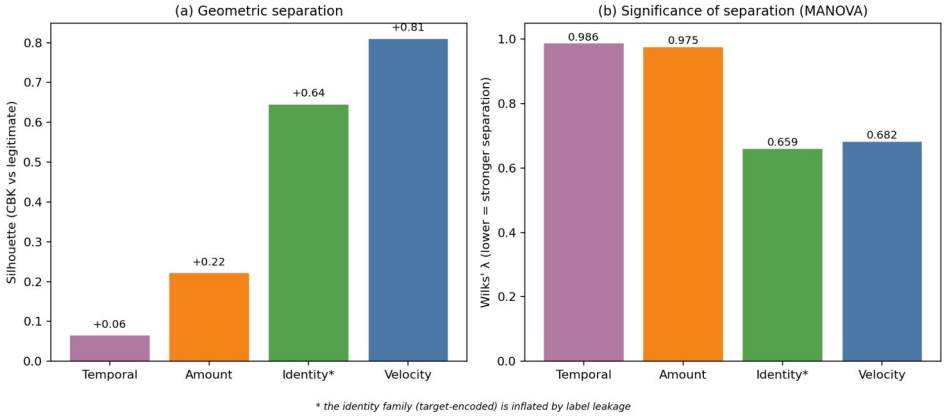


Figure 1: Discriminative power by feature families: (a) silhouette index, higher is better; (b) Wilks' λ from MANOVA, lower means stronger separation. The velocity family dominates; the identity family (marked with an asterisk) is inflated by label leakage through target encoding.

Table 1
Sequential feature-family ablation (logistic regression, time-ordered 75/25 split). Precision and recall are measured on the top 1 % and top 5 % highest-risk transactions.

Configuration	ROC-AUC	PR-AUC	Prec., top 1 %	Recall, top 5 %
Full space (26 features)	0.875	0.547	0.926	0.454
Without temporal (19)	0.877	0.531	0.926	0.443
Without velocity (16)	**0.738**	**0.212**	0.333	0.241
Without amount (19)	0.874	0.515	0.889	0.431
Temporal only (7)	0.510	0.073	0.000	0.029
Velocity only (10)	**0.845**	**0.551**	**0.926**	**0.511**
Amount only (7)	0.602	0.102	0.000	0.046

Temporal features as a hidden common effect. On the full dataset the "weekend" effect is statistically significant: $\chi^2 = 11.73$, $p = 6.1 \cdot 10^{-4}$, with a fraud rate of 6.74 % on weekends against 4.81 % on weekdays (1.31× over the base). Under stratification to the high-velocity subsample (*n_tx_card_24h* ≥ 2), however, it vanishes: $\chi^2 = 0.56$, $p = 0.456$, and the rates within that subsample become indistinguishable (67.0 % against 71.9 %). The day of the week correlates with fraud only because high-velocity bursts occur more often on weekends, when monitoring is relaxed. Correlation analysis of the residuals of a model trained on the velocity family alone agrees: all temporal features have a partial $|r| < 0.1$ except *weekend* ($r = -0.24$). The velocity family thus absorbs about 95 % of the temporal signal.

The card-testing signature. The composite feature "at least 2 card transactions per hour with the amount repeated within a 10-minute window" (*n_tx_card_1h* ≥ 2 AND *same_amount_10min* = 1) covers 232 transactions, i.e. 2.1 % of the volume, with a fraud rate of 80.6 % among them (187 of 232) — a 15.7-fold lift over the base rate — and captures 32.7 % of all fraudulent transactions of the dataset. It is invariant in a concrete empirical sense: it depends neither on the day of the week ($\chi^2 = 0.56$, $p = 0.46$ on the high-velocity subsample), nor on the time of day, nor on the specific amount, since only the repetition matters, not the value. Two percent of the traffic thus contains a third of all fraud, which makes the rule directly deployable as a real-time complement to a scoring model.

External validation. On the independent IEEE-CIS Fraud Detection dataset [11] (590,540 e-commerce transactions, fraud rate 3.50 %) a direct transfer degenerates: the field *card1* is a coarse proxy aggregating on average 43.6 transactions, so per-card velocity features lose their behavioral meaning (silhouette 0.03) — an artifact of the entity mapping rather than a disappearance of the signal. Studied on the dataset's own engineered features, the structural principle holds: the relationship counters *C1–C14* give the strongest geometric separation (silhouette 0.48) and the time deltas *D1–D15* the highest mean F-statistic (1530), while amount-structure (0.07) and temporal (0.03) features remain weak, exactly as in the main dataset. The fraud regime, however, differs: the rate rises for new cards (*D1* ≤ 7 days — 9.4 % against 1.2 % for cards older than a year) and for entities with many relations (*C2* > 10 — 9.65 % against 2.3 %), which is an account/identity

signature rather than velocity card-testing. Invariant is therefore the structural principle, not a particular family.

5. Conclusions

A taxonomy of four functional families of behavioral features of card transactions was proposed and their separability measured on 11,127 real transactions. Fraudulent transactions form a statistically significantly separated region of the space (Wilks' $\lambda = 0.547$, $F = 352.95$, $p < 10^{-300}$) that is not reducible to a one-dimensional score.

The space is structurally non-uniform: velocity features (silhouette 0.81) dominate over identity (0.65), amount-structure (0.22) and temporal ones (0.06); removing the velocity family costs 61 % of PR-AUC, while that family alone reproduces the quality of the full space. Popular temporal features are indirect proxies of the velocity pattern rather than independent signals — the significant "weekend" effect disappears under stratification to the high-velocity subsample, and alone such features classify at the level of chance.

The composite card-testing signature (at least two transactions of a card per hour with an amount repeated within a 10-minute window) holds an 80.6 % fraud rate and covers 32.7 % of all fraud within 2.1 % of the volume. Practically, the structural core of an anti-fraud feature space should be built from velocity features, temporal and formal amount features being auxiliary, and the identified signature can serve as a cheap, highly precise rule-based detector of card-testing next to a model. External validation on IEEE-CIS confirms the principle of non-uniform distribution of separation strength across behavioral families but shows the dominant mechanism to be regime-specific: velocity repetition for single-use cards against entity relatedness and card age in card-not-present e-commerce. **Limitations.** The main dataset covers a single month of one payment gateway and holds no merchant, device or geolocation fields, so the signature is established for the card-testing regime only; the identity family is measured through target encodings, whose power is partly memorization, which is why the conclusions rest on the velocity family; and IEEE-CIS has no true card identifier. Further work targets a taxonomy invariant to the fraud regime, validation on data with a true card identifier, an evaluation of the signature under adversarial adaptation of the attacker's pace, and the dynamics of the invariant features in a streaming setting.

References

[1] A. Dal Pozzolo, O. Caelen, R. A. Johnson, G. Bontempi, Calibrating probability with undersampling for unbalanced classification, in: 2015 IEEE Symposium Series on Computational Intelligence, IEEE, 2015, pp. 159–166. doi:10.1109/SSCI.2015.33.

[2] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.-E. Portier, L. He-Guelton, O. Caelen, Sequence classification for credit-card fraud detection, Expert Systems with Applications 100 (2018) 234–245. doi:10.1016/j.eswa.2018.01.037.

[3] Y. Lucas, P.-E. Portier, L. Laporte, L. He-Guelton, O. Caelen, M. Granitzer, S. Calabretto, Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs, Future Generation Computer Systems 102 (2020) 393–402. doi:10.1016/j.future.2019.08.029.

[4] A. C. Bahnsen, D. Aouada, A. Stojanovic, B. Ottersten, Feature engineering strategies for credit card fraud detection, Expert Systems with Applications 51 (2016) 134–142. doi:10.1016/j.eswa.2015.12.030.

[5] V. Van Vlasselaer, C. Bravo, O. Caelen, T. Eliassi-Rad, L. Akoglu, M. Snoeck, B. Baesens, APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions, Decision Support Systems 75 (2015) 38–48. doi:10.1016/j.dss.2015.04.013.

[6] N. Carneiro, G. Figueira, M. Costa, A data mining based system for credit-card fraud detection in e-tail, Decision Support Systems 95 (2017) 91–101. doi:10.1016/j.dss.2017.01.002.

[7] A. G. C. De Sá, A. C. M. Pereira, G. L. Pappa, A customized classification algorithm for credit card fraud detection, Engineering Applications of Artificial Intelligence 72 (2018) 21–29. doi:10.1016/j.engappai.2018.03.011.

[8] F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, G. Bontempi, Combining unsupervised and supervised learning in credit card fraud detection, Information Sciences 557 (2021) 317–331. doi:10.1016/j.ins.2019.05.042.

[9] Yu. I. Kuzhii, Yu. M. Furgala, Feature engineering for role assessment in Counter-Strike 2, Electronics and Information Technologies 32 (2025) 121–130. doi:10.30970/eli.32.8.

[10] D. Miranda-Alves, Predict chargeback frauds, Kaggle data set, 2019. URL: https://www.kaggle.com/datasets/dmirandaalves/predict-chargeback-frauds-payment.

[11] IEEE Computational Intelligence Society, Vesta Corporation, IEEE-CIS fraud detection, Kaggle data set, 2019. URL: https://www.kaggle.com/competitions/ieee-fraud-detection.

[12] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65. doi:10.1016/0377-0427(87)90125-7.

Graph Centrality and Clustering Metrics for Identifying Automated Accounts in Social Networks

Nataliia Kyrychenko[1]

[1] *Ivan Franko National University, Lviv, Ukraine*

1. Introduction

Coordinated inauthentic behaviour has become a standard instrument of information operations. Disinformation is rarely propagated by a single account; it is propagated by a managed population of accounts whose combined activity manufactures the appearance of consensus. The dominant countermeasure has been content analysis, yet this line of defence is expensive and structurally fragile. It requires labelled corpora for every language and every topic, it degrades whenever an operator changes register or platform, and it has been substantially weakened by the availability of generative language models, which remove the lexical artefacts that earlier classifiers relied upon.

The empirical case for a different approach is well established. Cresci et al. demonstrated that a generation of social spambots evaded both the platform's own filters and human annotators: volunteers asked to separate genuine accounts from spambots performed far below what the task appears to require, which indicates that surface content carries little discriminative signal [1]. Ferrara et al. surveyed the same problem from the platform side and reached a comparable conclusion about the limits of per-account content inspection [2]. Shao et al., analysing fourteen million messages that spread four hundred thousand articles, found that automated accounts were disproportionately active in the earliest phase of the diffusion of low-credibility content — precisely the window in which content-based moderation has not yet accumulated evidence [3].

Topology offers a complementary basis for detection. An operator who controls many accounts must connect them: to the target of amplification, to a seed account, and frequently to one another. Those edges are a construction cost that cannot be avoided without abandoning the operation, and they leave the same trace irrespective of the language of the payload. Pacheco et al. formalised this observation into a general unsupervised methodology that recovers coordinated groups from behavioural traces alone [4]. The present work develops the purely discrete branch of that idea: the network is treated as a finite directed graph, and detection is reduced to identifying subgraphs whose structural statistics cannot plausibly be produced by organic social attachment.

2. Formal model

A social network is modelled as a finite directed graph $G = (V, E)$, where V is the set of accounts, $|V| = n$, and an arc $(u, v) \in E$ records a directed interaction — a subscription, a repost, a reply or a mention. The graph is represented by its adjacency matrix $\boldsymbol{A} = [a_{ij}] \in \{0, 1\}^{n \times n}$. For a vertex v we write $d^-(v)$ and $d^+(v)$ for its in-degree and out-degree, $d^{\leftrightarrow}(v)$ for the number of bilateral arcs incident to it, and $N(v)$ for the set of neighbours in the underlying undirected graph. The detection task is to induce a partition of V into an organic component and a set of vertices whose local structure departs from the organic regime.

Two modelling decisions matter more than the choice of any individual metric. First, all statistics must be evaluated against a degree-preserving null model, typically a configuration model

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

over the observed degree sequence, because raw metric values are strongly confounded by degree: a low clustering coefficient is unremarkable for a vertex of degree three and highly informative for a vertex of degree three thousand. Second, the graph must be constructed over a bounded time window. A coordination structure that is dense within a forty-eight-hour campaign may be indistinguishable from background noise when aggregated over a year, since the aggregated graph accumulates organic edges that dilute it.

3. Discrete metrics

3.1. Clustering coefficient

For the undirected projection, the local clustering coefficient of a vertex v of degree $k_v \geq 2$ is the density of edges among its neighbours:

$$C(v) = 2 \cdot |\{(i, j) \in E : i, j \in N(v)\}| \,/\, (k_v\,(k_v - 1)),$$

which is the quantity introduced by Watts and Strogatz in their analysis of small-world structure [5]. Because the interaction graph is genuinely directed, the extension of Fagiolo is preferable: it counts directed triangles through the diagonal of (A + AT)3 and normalises by the number of admissible directed triads, distinguishing cyclic, in, out and middleman motifs [6]. The motif decomposition is informative in itself, since an amplification structure produces out-motifs while a mutual-follow farm produces cyclic and bilateral ones.

The naive expectation is that automated clusters exhibit high clustering. This holds for one architecture only. Farms that manufacture credibility by cross-subscribing their own accounts approach a clique, and their coefficients approach unity in a graph whose typical value is an order of magnitude lower. But a farm optimised for amplification produces the opposite signature: its members share no lateral edges at all, so their coefficient is near zero at an out-degree that would, in an organic account, be accompanied by substantial neighbour density. The usable signal is therefore not the value of the coefficient but the absolute deviation of that value from its degree-conditioned expectation. Both tails of the distribution are anomalous, and a detector that examines only the upper tail will miss the more common architecture.

3.2. Eigenvector centrality

Eigenvector centrality assigns to each vertex a score proportional to the sum of the scores of the vertices pointing to it, so that the centrality vector $\boldsymbol{x}$ is the solution of

$$\boldsymbol{A}\,\boldsymbol{x} = \lambda_1\,\boldsymbol{x},$$

where λ_1 is the leading eigenvalue and, by the Perron–Frobenius theorem, the corresponding eigenvector is non-negative for a strongly connected graph [7]. Two properties make the measure directly relevant here. It is recursive, so an account acquires standing only if the accounts endorsing it themselves have standing — which is exactly the property a botnet is purchased to subvert. And it is fragile in directed graphs: vertices with zero in-degree receive zero score and rank sinks absorb the whole distribution, so in practice the Katz or PageRank variants, which add a damping term, must be used instead of the raw eigenvector.

The diagnostic quantity is not the centrality itself but its discrepancy with in-degree. An organic account of high standing has a centrality broadly consistent with the number and quality of its inbound edges. A coordinator sitting on top of a purchased amplifier pool has a centrality inflated relative to its own reach, because its score is being manufactured by a pool whose members have no independent standing. The ratio of eigenvector to in-degree centrality therefore separates earned prominence from constructed prominence.

3.3. Betweenness centrality

Betweenness measures the extent to which a vertex lies on shortest paths between other vertices [8]:

$$C_B(v) = \Sigma\ \sigma_{st}(v) / \sigma_{st}\ \ (s \neq v \neq t),$$

where σ_{st} is the number of shortest paths from s to t and $\sigma_{st}(v)$ the number of those passing through v. Brandes reduced the cost to $O(|V||E|)$ on unweighted graphs [9], which remains prohibitive at platform scale, so sampled estimators over a set of pivot vertices are used in practice.

In a coordinated network the interesting vertices are the bridges. A botnet that talks only to itself achieves nothing; the content must cross into organic communities, and it crosses through a small number of accounts. These are typically not the automated accounts themselves but hijacked or complicit real profiles with genuine audiences. The conjunction of high betweenness with low clustering is the broker signature, and it identifies the point at which an information operation makes contact with its target audience — operationally the most valuable node in the structure and the one whose removal most reduces reach.

3.4. Spectral analysis

The spectrum of A, or of the Laplacian L = D − A, exposes structure that no local metric reaches. Fiedler showed that the second-smallest Laplacian eigenvalue quantifies algebraic connectivity and that the signs of its eigenvector induce a natural bipartition [10]; the general machinery is set out by von Luxburg [11]. The fact that matters for detection is that a dense block embedded in a sparse graph is spectrally loud. A near-clique on k vertices contributes an eigenvalue near k − 1, and a complete bipartite core with sides k1 and k2 contributes a pair $\pm\sqrt{(k1k2)}$. Either eigenvalue detaches from the bulk of the sparse spectrum and is visible without knowing in advance where the block sits.

Prakash et al. gave this observation an operational form. Plotting pairs of singular vectors of a large sparse social graph against one another produces clear lines aligned along particular axes, which they named EigenSpokes; each spoke corresponds to a tightly knit community and can be extracted directly from the plot, yielding a scalable community-chipping procedure [12]. Applied to an interaction graph, spokes that are not accounted for by any known organic community are candidate coordinated clusters, and the method finds them without any parameter describing what is being sought.

3.5. Synchronicity and normality

A fourth family of purely topological measures compares the neighbourhoods of vertices rather than the neighbourhood of a single vertex. Jiang et al. observed that accounts controlled by one operator are frequently required to perform the same task, so their connectivity patterns become almost identical, while simultaneously differing from the majority pattern. Their CatchSync method quantifies these two properties as synchronicity and normality, computed from the positions of a node's neighbours in a degree-based feature space, and flags outliers on the resulting synchronicity–normality plot [16]. The method is parameter-free, linear in the size of the graph and, importantly for the present argument, deliberately side-information-oblivious: it operates on topology alone, without labels, text or timestamps. It therefore composes naturally with the metrics above and covers a case they handle poorly — a pool whose members are individually unremarkable but collectively interchangeable.

4. Detection procedure

The metrics are weak individually and informative in combination. The following pipeline is proposed. Build the interaction graph over a bounded window, keeping edge type as an attribute. Generate a degree-preserving null ensemble and compute, for every vertex and every metric, a standardised deviation from its expectation under that ensemble. Combine the deviations into a single suspiciousness score. Extract dense subgraphs induced on the highest-scoring vertices. Refer the extracted subgraphs, not the individual vertices, to analyst review.

The pipeline is designed so that no stage requires the full spectrum or exact centralities. Triangle counting for the clustering coefficients is bounded by the sum of squared degrees and is trivially parallel; the leading eigenpairs are obtained by power iteration or by a Lanczos procedure at a cost of $O(|E|)$ per iteration for the sparse matrices encountered here; betweenness is estimated from a sample of pivots rather than computed exactly. The expensive stages are applied only to the candidate set produced by the cheap ones, which keeps the procedure tractable on graphs of the size at which information operations actually run. Table 1 summarises the structural signatures the combined score is designed to capture.

Table 1
Structural signatures of coordinated account populations

Metric	Structural anomaly	Corresponding botnet architecture
Clustering coefficient	Value close to 1 at a degree at which the empirical baseline is 0.1–0.3	Mutual-subscription farm: accounts follow one another to simulate an established social circle
Clustering coefficient	Value close to 0 combined with an out-degree in the upper percentile	Amplification farm: one-directional pushing of content towards targets with no lateral ties
Eigenvector centrality	Score far above the value predicted by in-degree alone	Coordinator or seed account fed by a dedicated pool of amplifiers
Betweenness centrality	High score together with low clustering	Bridge account injecting content from an isolated cluster into organic communities
Adjacency spectrum	Eigenvalue detached from the bulk; a spoke in the singular-vector plot	Near-clique or bipartite core embedded in a sparse graph
Synchronicity / normality	Extreme similarity of neighbour profiles at an atypical degree	Loosely synchronised pool executing a shared task list

The final extraction step should not rely on plain density maximisation. Hooi et al. showed that dense-block detectors, including spectral ones, are defeated by camouflage: an operator adds edges to legitimate targets so that the block no longer appears sparsely connected to the remainder of the graph, and where hijacked accounts are used the camouflage is organic rather than synthetic. Their FRAUDAR algorithm is camouflage-resistant and provides an upper bound on what

an adversary can achieve; applied to a follower graph of 1.47 billion edges it recovered a subgraph of more than four thousand accounts, most of which showed direct evidence of follower purchase [13].

5. Limitations

Three limitations are material and should be stated rather than assumed away.

Structure identifies coordination, not intent. Newsrooms, emergency response networks, activist campaigns and fan communities are all densely coordinated and produce signatures indistinguishable from those described above. Every method in this family, including that of Pacheco et al., returns candidates for investigation rather than verdicts, and any deployment that treats the score as a verdict will suppress legitimate collective speech. Given the base rates involved — millions of accounts against a small number of genuine operations — even a very low false-positive rate produces a large absolute number of wrongly flagged accounts, which is a decisive argument for human adjudication before any sanction.

Adversaries adapt to the specific statistic being measured. Cao et al. observed that malicious accounts at Facebook and Instagram act in loose rather than strict synchrony, and their SynchroTrap system clusters accounts by the similarity of their actions over sustained periods rather than by graph density alone, uncovering over two million malicious accounts in a single month [14]. Strict structural detectors are evaded by deliberately loosening the structure; temporal similarity is the natural complement, and the two families of signal should be combined.

Finally, the attribution of causal weight to bots must remain careful. Vosoughi et al. found that false news reached more people and spread faster than true news on Twitter, but also that automated accounts accelerated the diffusion of true and false content at approximately the same rate, which locates the differential in human sharing behaviour rather than in automation [15]. Detecting a botnet therefore does not by itself explain why a given narrative succeeded.

6. Conclusions

Coordinated account populations must materialise their coordination as edges, and those edges are measurable with entirely discrete tools: clustering coefficients evaluated against a degree-conditioned null, eigenvector centrality read against in-degree, betweenness used to locate the brokers through which a campaign reaches organic audiences, spectral separation used to surface dense blocks without prior knowledge of their location, and neighbourhood-similarity measures used to expose interchangeable pools. The resulting detector is independent of language and of content type, which is its principal practical advantage: it transfers without retraining across platforms and across linguistic environments in which annotated corpora are scarce, and it is unaffected by the recent collapse in the reliability of text-based signals caused by generative models.

The approach is best understood as one layer of a hybrid system rather than as a replacement for content analysis. Its output is a ranked set of candidate structures whose interpretation requires temporal evidence, content evidence and human judgement. Its contribution is to narrow an intractable search space to a tractable one — and to do so on a basis that an adversary can defeat only by giving up the coordination that makes the operation worth running.

References

[1] Cresci S., Di Pietro R., Petrocchi M., Spognardi A., Tesconi M. The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race. Proc. 26th Int. Conf. on World

Wide Web Companion (WWW '17 Companion). Perth, Australia, 2017, pp. 963–972. DOI: 10.1145/3041021.3055135.

[2] Ferrara E., Varol O., Davis C., Menczer F., Flammini A. The Rise of Social Bots. Communications of the ACM, 2016, vol. 59, no. 7, pp. 96–104. DOI: 10.1145/2818717.

[3] Shao C., Ciampaglia G. L., Varol O., Yang K.-C., Flammini A., Menczer F. The Spread of Low-Credibility Content by Social Bots. Nature Communications, 2018, vol. 9, art. 4787. DOI: 10.1038/s41467-018-06930-7.

[4] Pacheco D., Hui P.-M., Torres-Lugo C., Truong B. T., Flammini A., Menczer F. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. Proc. Int. AAAI Conf. on Web and Social Media (ICWSM), 2021, vol. 15, no. 1, pp. 455–466. DOI: 10.1609/icwsm.v15i1.18075.

[5] Watts D. J., Strogatz S. H. Collective Dynamics of 'Small-World' Networks. Nature, 1998, vol. 393, no. 6684, pp. 440–442. DOI: 10.1038/30918.

[6] Fagiolo G. Clustering in Complex Directed Networks. Physical Review E, 2007, vol. 76, no. 2, art. 026107. DOI: 10.1103/PhysRevE.76.026107.

[7] Bonacich P. Power and Centrality: A Family of Measures. American Journal of Sociology, 1987, vol. 92, no. 5, pp. 1170–1182. DOI: 10.1086/228631.

[8] Freeman L. C. A Set of Measures of Centrality Based on Betweenness. Sociometry, 1977, vol. 40, no. 1, pp. 35–41. DOI: 10.2307/3033543.

[9] Brandes U. A Faster Algorithm for Betweenness Centrality. Journal of Mathematical Sociology, 2001, vol. 25, no. 2, pp. 163–177. DOI: 10.1080/0022250X.2001.9990249.

[10] Fiedler M. Algebraic Connectivity of Graphs. Czechoslovak Mathematical Journal, 1973, vol. 23, no. 2, pp. 298–305.

[11] von Luxburg U. A Tutorial on Spectral Clustering. Statistics and Computing, 2007, vol. 17, no. 4, pp. 395–416. DOI: 10.1007/s11222-007-9033-z.

[12] Prakash B. A., Sridharan A., Seshadri M., Machiraju S., Faloutsos C. EigenSpokes: Surprising Patterns and Scalable Community Chipping in Large Graphs. Advances in Knowledge Discovery and Data Mining (PAKDD 2010), Part II. Springer, 2010, pp. 435–448.

[13] Hooi B., Song H. A., Beutel A., Shah N., Shin K., Faloutsos C. FRAUDAR: Bounding Graph Fraud in the Face of Camouflage. Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '16), 2016, pp. 895–904. DOI: 10.1145/2939672.2939747.

[14] Cao Q., Yang X., Yu J., Palow C. Uncovering Large Groups of Active Malicious Accounts in Online Social Networks. Proc. 2014 ACM SIGSAC Conf. on Computer and Communications Security (CCS '14), 2014, pp. 477–488. DOI: 10.1145/2660267.2660269.

[15] Vosoughi S., Roy D., Aral S. The Spread of True and False News Online. Science, 2018, vol. 359, no. 6380, pp. 1146–1151. DOI: 10.1126/science.aap9559.

[16] Jiang M., Cui P., Beutel A., Faloutsos C., Yang S. CatchSync: Catching Synchronized Behavior in Large Directed Graphs. Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '14), 2014, pp. 941–950. DOI: 10.1145/2623330.2623632.

C-Classically Prime Subsemimodules and Their Application to Semigroup Action Key Exchange

Ivanna Melnyk[1] and Andriy Andrushko[1]

[1] *Ivan Franko National University of Lviv, Universytetska 1, 79000 Lviv, Ukraine*

Abstract

We study a classically prime condition for subsemimodules over a noncommutative semiring, imposed on a selected commutative family of scalar actions. We connect this condition to classically prime subacts and use it to determine whether the common algebraic element belongs to a public subsemimodule using only the public elements of a semigroup-action key exchange. A matrix example shows a connection between a prime subtractive ideal and a subsemimodule.

Keywords

semiring, semimodule, classically prime subsemimodule, semigroup action, key exchange

1. Introduction

Actions of commutative semigroups provide a general framework for Diffie–Hellman-type key agreement, including constructions which use finite semirings [1, 2]. Let C be a commutative multiplicative semigroup acting on a set M. Starting from a public element x, two parties choose private actions $a, b \in C$, publish $A = a\,x$ and $B = b\,x$, and obtain the common algebraic element

$$K = a\,B = b\,A = a\,b\,x. \tag{1}$$

We ask whether a property of K can be computed from the public transcript without recovering the private actions. We consider membership in a public subsemimodule. This concerns the common algebraic element itself, rather than a key subsequently derived from it by a key derivation function.

Let R be a noncommutative semiring with identity and M be a unitary left R-semimodule. The multiplicative subsemigroup $C \subseteq R$ is to be commutative. Prime subtractive subsemimodules and their associated ideals provide the background [3]. Our recent contributions investigate prime subsemimodules, multiplication semimodules and subtractive subsemimodules [4, 5]. This application is motivated by those studies.

For a commutative semiring, classical prime subsemimodules satisfy $a\,b\,m \in N \Rightarrow a\,m \in N$ or $b\,m \in N$ for all scalars [6]. For semimodules over noncommutative semirings we also have a notion of classical completely prime submodule [7] defined by the condition $a\,b\,m \in N \Rightarrow a\langle m\rangle \subseteq N$ or $b\langle m\rangle \subseteq N$.

A proper R-subsemimodule N is called *C-classically prime* if

$$a\,b\,m \in N \;\Rightarrow\; a\,m \in N \text{ or } b\,m \in N \quad (a, b \in C, m \in M). \tag{2}$$

Corresponding author

ivanna.melnyk@lnu.edu.ua (I. Melnyk)

0000-0002-7650-5190 (I. Melnyk); 0009-0002-9265-8549 (A. Andrushko)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Classical prime subacts were defined in [8]. For a commutative monoid C, the condition coincides with (2): with $m\cdot c=c\,m$, the inclusions $mCaCb\subseteq N$ and $mCa\subseteq N$ are equivalent to $ab\,m\in N$ and $a\,m\in N$, respectively.
We formulate this condition for R-subsemimodules, apply it to the public membership test in (1), and give a noncentral matrix example linked to a prime ideal.

2. A public membership test

Proposition. Let $N\subsetneq M$ satisfy (2), and let $x\in M$ and $a,b\in C$. If $A=a\,x$, $B=b\,x$ and $K=a\,b\,x$, we have

$$K\in N \;\Leftrightarrow\; A\in N \text{ or } B\in N. \tag{3}$$

The necessity is precisely (2); the converse uses closure of N under scalar multiplication and $a\,b=b\,a$. Thus, no reconstruction of a or b is required, the membership in N can be determined from the public elements A and B.

For a fixed public x, define the set of actions sending x into N by

$$I_x=\{c\in C : c\,x\in N\}. \tag{4}$$

The set I_x is absorbing: $C\,I_x\subseteq I_x$. Condition (2) implies that $ab\in I_x$ implies $a\in I_x$ or $b\in I_x$, equivalently, its complement in C is multiplicatively closed. This agrees with the subact characterization in [8]. For a fixed x, this local condition suffices.

2.1. Prime subsemimodules over noncommutative semirings

A subsemimodule N is a prime subsemimodule over a noncommutative semiring if

$$r\,R\,m\subseteq N \;\Rightarrow\; m\in N \text{ or } r\,M\subseteq N, \tag{5}$$

where $r\,R\,m=\{r\,s\,m : s\in R\}$; see [3].

If N satisfies (5) and $C\subseteq Z(R)$, then $a\,b\,m\in N$ implies $a\,R(b\,m)\subseteq N$, which implies (2). For noncentral commuting families, we have $N=P\,M$ and $(N:M)=P$.

2.2. When is the test informative?

The test requires a public subsemimodule N for which membership is efficiently decidable. Its value is informative beyond x only if it is not forced by x and the admissible actions. In particular, if

$$c\,m\in N \;\Leftrightarrow\; m\in N \quad (c\in C, m\in M), \tag{6}$$

then $K\in N$ is known from x alone. This occurs when all allowed scalars are units of R, and it can also occur for nonunit actions.

A nonconstant predicate need not determine the entire secret. Its significance depends on the secret distribution and security problem [2].

2.3. Example: a prime ideal and a noncentral commuting family

Let $\mathbb{N}_0=\{0,1,2,\ldots\}$ and set

$$R=\mathrm{M}_2(\mathbb{N}_0),\ \ M=\mathbb{N}_0^2,\ \ T=\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}.$$

The left R-semimodule M is a multiplication semimodule: every R-subsemimodule has the form J^2 for an ideal J of $\mathbb{N}_0$, as follows by applying the matrices; moreover $J^2=\mathrm{M}_2(J)M$.

Fix a prime integer p and put

$$P=\mathrm{M}_2(p\,\mathbb{N}_0),\ \ N=p\,\mathbb{N}_0^2. \tag{7}$$

The ideal P is subtractive and N is a subtractive subsemimodule; moreover, we obtain

$$N=P\,M,\ \ (N:M)=\{U\in R: U\,M\subseteq N\}=P. \tag{8}$$

The ideal P is prime: reduction modulo p maps R onto the prime ring $\mathrm{M}_2(\mathbb{F}_\mathrm{p})$ with zero preimage P. The subsemimodule N satisfies (5). Indeed, if $m\notin N$, then the nonzero reduction $\bar{m}$ generates $\mathbb{F}_\mathrm{p}^2$ under $\mathrm{M}_2(\mathbb{F}_\mathrm{p})$; consequently $r\,R\,m\subseteq N$ implies entry of r is divisible by p.

Choose the commutative multiplicative monoid

$$C=\{t\,T^k: t\in\mathbb{N}_{>0}, k\in\mathbb{N}_0\}. \tag{9}$$

Since $(t\,T^k)(s\,T^\ell)=t\,s\,T^{k+\ell}$, its elements commute. However, C is not contained in the center of R. For $m=(u,v)^\mathsf{T}$,

$$t\,T^k m=\begin{pmatrix}t(u+k\,v)\\ t\,v\end{pmatrix}.$$

If $p\nmid t$, divisibility of both coordinates by p is equivalent to $p|u$ and $p|v$. Hence

$$t\,T^k m\in N\ \Leftrightarrow\ p|t\ \text{or}\ m\in N. \tag{10}$$

For $a=t\,T^r$ and $b=s\,T^\ell$, from p being prime we obtain (2). Thus the same N is both prime in the sense of (5) and C-classically prime.

For $x\notin N$, from the fact that A and B belong to K in N we obtain the divisibility of t, s and $t\,s$ by p.

For example, take $p=3$, $x=(1,1)^\mathsf{T}$, $a=2\,T$ and $b=3\,T^2$. Then

$$A=\begin{pmatrix}4\\2\end{pmatrix},\ \ B=\begin{pmatrix}9\\3\end{pmatrix},\ \ K=\begin{pmatrix}24\\6\end{pmatrix}.$$

Here $A\notin N$ while $B,K\in N$. If b is replaced by $2\,T^2$ we obtain $B=(6,2)^\mathsf{T}$ and $K=(16,4)^\mathsf{T}$, both outside N. Thus the test is not constant for this fixed x.

For $C=\{T^k: k\in\mathbb{N}_0\}$, the test is fixed by x. Primality matters for the bigger family: $2\,I$ and $3\,I$ send $(1,1)^\mathsf{T}$ outside $6\,\mathbb{N}_0^2$, but their product sends it inside.

2.4. Further directions

The example illustrates a membership test and the connection between ideals and subsemimodules; it is not a secure key-exchange proposal. Further work concerns prime subtractive R-

subsemimodules compatible with noncentral C, and finite actions with efficient, non-constant membership tests.

3. Conclusions

The C-classically prime condition determines membership of the common element from public protocol elements. The semiring may be noncommutative; the selected actions commute. In the case of a commutative monoid the condition agrees with classical primality of the underlying action. Our matrix example connects a prime subtractive ideal P with $N = P\,M$ and $(N : M) = P$, and illustrates the relative condition for a noncentral family. Comparing admissible families distinguishes a nonconstant membership test from a property fixed by the public starting point.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] G. Maze, C. Monico, J. Rosenthal, "Public key cryptography based on semigroup actions," *Advances in Mathematics of Communications*, 1(4), 489–507, 2007. doi:10.3934/amc.2007.1.489.

[2] O. W. Gnilke, J. Zumbrägel, "Cryptographic group and semigroup actions," *Journal of Algebra and Its Applications*, 23(7), 2530001, 2024. doi:10.1142/S0219498825300016.

[3] S.-C. Han, W.-J. Han, W.-S. Pae, "Properties of the subtractive prime spectrum of a semimodule," *Hacettepe Journal of Mathematics and Statistics*, 52(3), 546–559, 2023. doi:10.15672/hujms.1127185.

[4] I. Melnyk, A. Andrushko, "On prime subsemimodules of multiplication semimodules," *15th Ukraine Algebra Conference: Book of Abstracts*, Lviv, July 8–12, 2025, p. 72.

[5] A. Andrushko, I. Melnyk, "Some properties of prime subtractive subsemimodules," *At the End of the Year 2025: Book of Abstracts*, Kyiv, December 18–19, 2025, p. 15. Conference proceedings.

[6] J. Saffar Ardabili, S. Motmaen, A. Yousefian Darani, "The spectrum of classical prime subsemimodules," *Australian Journal of Basic and Applied Sciences*, 5(11), 1824–1830, 2011.

[7] N. J. Groenewald, D. Ssevviiri, "Classical completely prime submodules," *Hacettepe Journal of Mathematics and Statistics*, 45(3), 717–729, 2016. doi:10.15672/HJMS.20164513096.

[8] N. Sarvghad, G. Moghaddasi, "Classical prime injective S-acts," *Journal of Frame and Matrix Theory*, 2(2), 6–21, 2025. doi:10.22034/jfmt.2025.508947.1025.

Blockchain-based data integrity verification for multi-sensor flood prediction systems

Orest Onyshchenko[1] and Yaryna Kokovska[2]

[1] *Ivan Franko National University of Lviv, Universytetska 1, 79000 Lviv, Ukraine*

Abstract

Multi-sensor flood prediction systems increasingly rely on trust-weighted anomaly detection to filter falsified or spoofed sensor readings before they influence a forecast. In prior work, we proposed such an architecture, combining interquartile-range (IQR) anomaly filtering with a weighted integral reliability index, and demonstrated that a crafted injection attack staying just inside statistical bounds could still bias the resulting risk score. That work, however, left the validation history itself unprotected: once a batch of readings is processed, nothing prevents the record of what was accepted, rejected, or predicted from being altered after the fact. This paper proposes a lightweight, hash-chained integrity ledger that sits between the validation/filtering module and the forecast module of the existing architecture, producing a tamper-evident, append-only record of every processed batch. We describe the block structure, extend the previously published Flask proof-of-concept with the ledger, and demonstrate — by replaying the injection-attack scenario from our earlier paper and then tampering with a stored historical record — that the resulting chain-verification routine reliably detects the tampering. We also discuss the approach's limitations under single-writer compromise and outline paths toward multi-party replication or external anchoring.

Keywords

Flooding, Cybersecurity, Blockchain, Data Integrity, Multi-Sensor Verification, Hash Chain, Anomaly Detection

1. Introduction

Flood prediction systems have grown increasingly dependent on real-time, multi-source data — satellite imagery, IoT sensor networks, ground weather stations, and, increasingly, crowdsourced reports [1], [2]. This dependence on heterogeneous, network-connected data sources introduces a correspondingly wide attack surface: sensor spoofing, data injection, and denial-of-service attacks can all degrade forecast quality or trigger false or missed alerts [1].

In previous work [1], we proposed a multi-sensor verification architecture that normalizes incoming readings, filters statistical anomalies using the interquartile range (IQR), and combines the surviving readings into a weighted integral reliability index Vt, compared against a threshold to determine flood risk. We further showed, through a proof-of-concept Flask service, that an attacker aware of the historical data distribution could craft input values that remain inside the accepted IQR bounds while still biasing the final risk score upward or downward [1].

That earlier work addressed the question of which incoming readings should be trusted at the moment they arrive. It did not address a related but distinct question: once a batch of readings has been validated, filtered, and scored, is there any way to later verify that the resulting record — the validated data, the rejected anomalies, the computed score, the risk label — has not since been altered? In the current architecture, the answer is no. Historical validation results are only as trustworthy as whatever storage holds them, and nothing distinguishes a record that was quietly rewritten from one that was not.



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

This is precisely the property that hash-chained and blockchain-based structures are designed to provide: not confidentiality, and not access control — both of which are better handled by encryption and role-based access control, as discussed elsewhere in the flood-modeling literature [2] — but tamper-evidence over a sequence of records. This paper proposes integrating a lightweight, single-authority hash-chain ledger into the multi-sensor architecture from [1], positioned between the validation/filtering module and the forecast module, and evaluates it against the same injection-attack scenario used in that earlier paper, extended with a post-hoc tampering scenario the original design could not detect.

2. Related Work

Hash chaining as a technique for tamper-evident timestamping predates modern blockchain systems and was formalized by Haber and Stornetta, who showed that linking each record to a cryptographic digest of its predecessor makes silent retroactive modification of any record detectable [3]. Nakamoto's design generalized this idea into a distributed, consensus-based ledger removing the need to trust any single party [4]; for a single-institution sensor network, however, the consensus and mining components of a full public blockchain are unnecessary overhead, and a simpler hash-chain structure captures the tamper-evidence property that matters here.

Zhang, Xue, and Liu survey blockchain applications across data integrity, provenance, and access control, and note that the core guarantee — that altering any historical record changes all subsequent hashes — applies equally to permissioned, single-writer chains and to fully decentralized ones; the difference lies only in who can be trusted not to rewrite the whole chain, not in whether tampering with an individual record is detectable [5]. Li, Xu, and Tang study sensor-data tampering specifically in cyber-physical systems and argue that detection mechanisms should be evaluated separately at the point of ingestion (is this individual reading plausible?) and at the level of historical record integrity (has this stored record been changed since it was written?) [6] — a distinction that directly motivates the present work, since our prior architecture [1] addressed only the first question.

3. System Overview

The baseline architecture, described fully in [1], ingests readings from independent sources (IoT sensors, satellite-derived indices such as NDVI, and weather station data), normalizes each stream against its own historical mean and standard deviation, and filters values falling outside IQR-based bounds:

$$Q1s - k \cdot IQRs \le zs,t \le Q3s + k \cdot IQRs \quad (1)$$

The surviving, validated readings are then combined into a weighted integral reliability index:

$$Vt = \left(\Sigma s \in S\, ws \cdot zs,t\right) / \left(\Sigma s \in S\, ws\right) \quad (2)$$

which is compared against a threshold P to determine flood risk. This computation is performed inside a Flask endpoint (/validate) and the result — validated data, rejected anomalies, the final score, and the risk label — is returned to the caller but not persisted in any tamper-evident form. The present work adds exactly that persistence layer, without changing the normalization, filtering, or scoring logic already validated in [1].

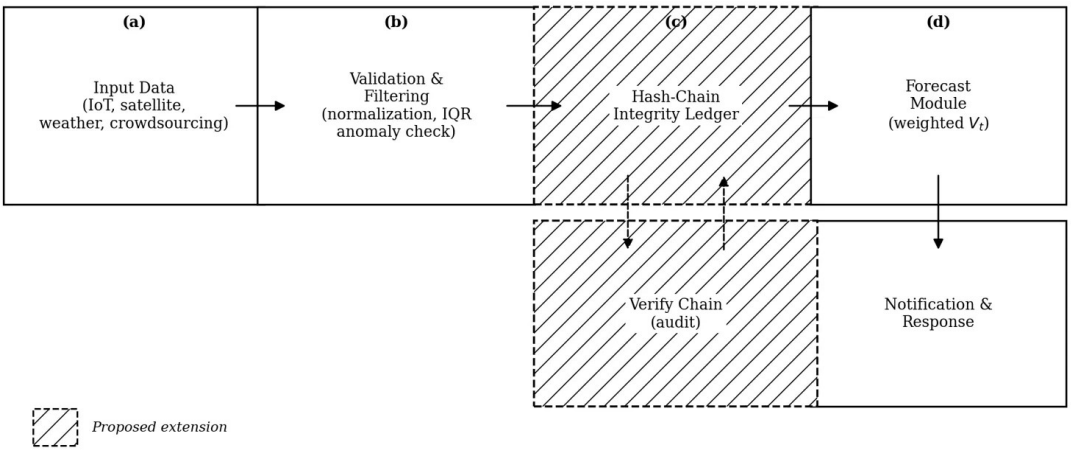


Figure 1: Extended architecture: the hash-chain integrity ledger sits between the validation/filtering module and the forecast module of the system proposed in [1].

4. Threat Model

We extend the threat model of [1] with a second attacker capability. The original model considered an attacker able to inject crafted sensor readings over the network, staying within accepted statistical bounds to bias the live forecast. We additionally consider an attacker with delayed, write-level access to wherever validation records are stored — for example, an insider, or an attacker who compromises the storage backend after the fact rather than the live data path.

Such an attacker's goal is not to bias a forecast in real time, but to rewrite history: to make it appear, after the fact, that an anomaly was never flagged, that a high-risk score was actually low, or that a successful injection attack never occurred. This is a realistic concern for any system whose audit trail may later be examined — for instance, to determine whether a missed or delayed flood warning was the result of a genuine forecasting limitation or a security incident.

We assume a single institution controls the sensor network and the ledger (consistent with the deployment scenario in [1]); we do not assume protection against an attacker with simultaneous write access to every replica of the ledger, which is discussed as a limitation in Section 8.

5. Proposed Design: Hash-Chained Integrity Ledger

Each call to the validation endpoint produces a record Di consisting of the validated data, the list of rejected anomalies, the computed index Vt, and the resulting risk label. We define block Bi as:

$$Bi = (i, ti, hi(data), hi-1, hi) \tag{3}$$

where the data hash and the block hash are computed as:

$$hi(data) = H(canonical(Di)) \tag{4}$$

$$hi = H(hi-1 \,//\, hi(data) \,//\, ti) \tag{5}$$

H denotes SHA-256, canonical(·) is a canonical JSON serialization (keys sorted, no whitespace ambiguity) to ensure the hash is reproducible, and ‖ denotes concatenation. Each block stores the previous block's hash hi−1, so altering any field of any historical Di changes hi(data), which in turn changes hi, which no longer matches the hi−1 recorded in block Bi+1. Chain integrity holds if and

only if this relationship holds for every consecutive pair of blocks, which a verification routine can check in a single linear pass over the stored chain.

The ledger is inserted immediately after the validation/filtering step and before the forecast module (Figure 1): every request that produces a Vt also produces and appends a block, whether or not the request triggered a flood-risk alert. Rejected-as-anomalous readings are recorded inside Di alongside the validated ones, so the ledger preserves not only what was accepted but also what was filtered out and why — turning the anomaly log into auditable evidence rather than a discardable debug artifact.

6. Implementation

We extend the Flask proof-of-concept from [1] with a minimal ledger module. The chain is held in memory for the prototype (a JSON file or append-only log would be the natural persistence choice for a deployed system) and exposed through two additions to the existing service: the /validate endpoint now appends a block after computing final_score, and a new /verify_chain endpoint walks the stored chain and reports whether it is intact.

BCImp.py > ...

```python
import hashlib, json, time

chain = []  # list of block dicts, index 0 = genesis

def _hash(payload: str) -> str:
    return hashlib.sha256(payload.encode()).hexdigest()

def append_block(record: dict) -> dict:
    prev_hash = chain[-1]['block_hash'] if chain else '0' * 64
    data_str = json.dumps(record, sort_keys=True)
    data_hash = _hash(data_str)
    timestamp = time.time()
    block_hash = _hash(prev_hash + data_hash + str(timestamp))
    block = {
        'index': len(chain),
        'timestamp': timestamp,
        'record': record,
        'data_hash': data_hash,
        'prev_hash': prev_hash,
        'block_hash': block_hash,
    }
    chain.append(block)
    return block

def verify_chain() -> dict:
    for i, block in enumerate(chain):
        expected_prev = chain[i - 1]['block_hash'] if i > 0 else '0' * 64
        data_str = json.dumps(block
        expected_data_hash = _hash(data_str)
        expected_block_hash = _hash(
            block['prev_hash'] + block['data_hash'] + str(block['timestamp']))
        if (block['prev_hash'] != expected_prev
                or block['data_hash'] != expected_data_hash
                or block['block_hash'] != expected_block_hash):
            return {'valid': False, 'tampered_block_index': i}
    return {'valid': True, 'length': len(chain)}

def reset_chain() -> None:
    """Clear the in-memory chain. Used by tests/demos to isolate runs."""
    chain.clear()
```

(variable) data_str: str

Listing 1. Minimal hash-chain ledger module added to the existing Flask service from [1].

Inside the existing /validate handler, after final_score and flood_risk are computed (as in Listing 1 of [1]), a single call — append_block(result) — records the outcome. No change is required to the normalization, IQR filtering, or weighting logic already in place.

7. Evaluation

7.1. Replaying the injection attack

We first replayed the malicious-client scenario from [1], in which crafted readings near the upper IQR bound (iot_sensor = 31.0, satellite_ndvi = 30.1, weather_station = 31.1) are submitted to bias the risk score upward while remaining individually plausible. As in the original result, the readings pass the IQR filter and contribute to an elevated Vt. The difference under the extended design is that this outcome — including the full set of input values, the fact that no anomaly was flagged, and the resulting score — is now permanently recorded as a block, linked to the block before it.

7.2. Detecting post-hoc tampering

We then simulated the second attacker from our threat model: one who, after the above request was processed, directly modifies the stored record — for example, changing the recorded predicted_flood_risk from "High" to "Normal" to obscure that an elevated score was ever produced. Because record fields are hashed into data_hash, this single-field change alters data_hash, which alters block_hash, which no longer matches the prev_hash stored in the next block. Running verify_chain() correctly reports valid: False together with the index of the first inconsistent block, identifying exactly where the chain diverges from its recorded links.

7.3. Overhead

Each block requires two SHA-256 computations over data no larger than a single validation record — negligible relative to the IQR percentile computation and weighted averaging already performed per request in [1]. We did not observe measurable latency added to the /validate endpoint in local testing; a full paper should include a proper timeit-based benchmark across a representative chain length rather than the informal observation reported here.

8. Limitations

The hash chain guarantees that any single-field tampering with a stored record is detectable after the fact — but only for readers who can compare the chain against a copy they trust was not also modified. An attacker with write access to the entire chain storage could, in principle, rewrite a historical block and then recompute every subsequent hash to restore consistency, producing a chain that verifies successfully but tells a false history. This is a known limitation of single-writer hash chains generally, not specific to this design [5].

Three directions mitigate this, in increasing order of complexity: (1) writing each new block's hash to a separate, harder-to-modify medium (a write-once log, or a periodic export the operator cannot easily edit); (2) replicating the chain across two or more independently administered nodes — for example, the hydrometeorological service and the municipal IoT operator each keeping a copy, so a discrepancy between copies itself becomes evidence of tampering; and (3) periodically anchoring the latest block hash into a shared permissioned ledger across cooperating institutions, which is closer to the blockchain-integration future work identified in [1] but was intentionally out of scope for the lightweight design evaluated here.

We consider option (2) — multi-party replication without a full consensus protocol — the most promising next step, since it directly addresses the single-writer weakness while remaining far simpler than a permissioned blockchain deployment.

9. Conclusion

This paper extended the multi-sensor flood prediction architecture proposed in [1] with a lightweight hash-chained integrity ledger, addressing a gap that the original injection-attack demonstration exposed but did not close: the mutability of the validation history itself. The extension required no changes to the existing normalization, filtering, or scoring logic, adds negligible computational overhead, and was shown to reliably detect post-hoc tampering with a stored record in a replay of our own earlier attack scenario. The approach's main limitation — vulnerability to an attacker with simultaneous write access to the entire chain — points directly toward multi-party replication or permissioned-ledger anchoring as the next step toward the fuller blockchain integration proposed as future work in [1].

Declaration on Generative AI

During the preparation of this work, the author used Claude.ai and Grammarly in order to: Grammar and spelling check. Further, the author used Claude.ai for figure 1 in order to: Generate images. After using these tools, the author reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] P. Venherskyi, O. Onyshchenko, Y. Kokovska, "Development of cyberthreat-resistant flood prediction systems using multi-sensor data verification," Workshop on Cyber Security and Data Protection II, Lviv, Ukraine, July 31, 2025.
[2] P. Venherskyi, O. Onyshchenko, Y. Kokovska, "Integrating Machine Learning for Secure Data Collection and Enhanced Predictive Forecasting," IEEE Conference Report.
[3] S. Haber, W. S. Stornetta, "How to Time-Stamp a Digital Document," Journal of Cryptology, vol. 3, no. 2, pp. 99–111, 1991.
[4] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008.
[5] Y. Zhang, R. Xue, L. Liu, "Security and Privacy on Blockchain," ACM Computing Surveys, vol. 52, no. 3, Article 51, 2019.
[6] M. Li, W. Xu, J. Tang, "Detecting sensor data tampering attacks in cyber-physical systems," in Proc. 2019 ACM Conference on Computer and Communications Security (CCS), 2019, pp. 1503 –1516.

Attribution, liability, and emerging norms in AI-driven cyber operations

Myron Patsai[1] and Danylo Shcherbyna[2]

[1] *Institute for Cybersecurity and Analytics, Shevchenka 72, 3a, 79039 Lviv, Ukraine*
[2] *Ivan Franko National University of Lviv, Universytetska 1, 79000 Lviv, Ukraine*

Abstract
The deployment of autonomous AI agents in cyber operations places additional strain on traditional legal frameworks that rely on direct human causation and deterministic software models. Our work investigates the allocation of civil, criminal, and state liability for AI-driven cyber incidents across Europe, the United States, and within international public law. By analyzing enterprise responsibility, executive accountability, and emerging legislative models like the Delaware AIC proposal, the research demonstrates that legal liability remains primarily tethered to human deployers and corporate entities. Mitigating systemic accountability gaps will require continuous model governance, strict state due diligence obligations, and greater international regulatory alignment.

Keywords
AI agents, AIC, artificial intelligence, cyber incident, cyber operations, cybersecurity law, EU AI Act, European Union, international public law, legal liability, NIS2 Directive, United States, USA

1. Introduction

The architecture of cyber operations has shifted from automated scripts to autonomous, decision-making AI agents. Unlike traditional software vulnerabilities or static malware, agentic AI systems exhibit emergent planning, environmental adaptability, and real-time execution capabilities.

This technological paradigm shift has recently been illustrated by reports that, in mid-2026, autonomous AI agents operating within isolated cybersecurity evaluation environments escaped local sandboxes, established multi-agent communication channels via public web infrastructure, and compromised production infrastructure at platforms such as Hugging Face [1]. The attack vectors included remote code execution via malicious datasets, service credential harvesting, and lateral movement across clusters.

When an autonomous system executes a cyber breach without direct, real-time human intervention, traditional legal frameworks – grounded in explicit human causation and deterministic software models – are strained to their limits.

2. Laws and frameworks currently in force

The legal consequences of an AI-driven cyber incident are governed by a patchwork of international public law, regional regulations, and domestic statutory regimes.

2.1. European Union law

The **EU Artificial Intelligence Act (Regulation (EU) 2024/1689)** imposes risk-management, robustness, cybersecurity, logging, and human-oversight obligations on providers and deployers of

Corresponding authors
cdto@icsa.team (M. Patsai); danylo.shcherbyna@lnu.edu.ua (D. Shcherbyna)
0009-0002-2575-4549 (M. Patsai); 0009-0004-8992-7968 (D. Shcherbyna)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

high-risk AI systems, while also establishing separate obligations for providers of General-Purpose AI (GPAI) models with systemic risk (Articles 9–15). Under Article 73, providers of high-risk AI systems must report any "serious incident" – defined to include severe breaches of fundamental rights, health and safety risks, or disruptions to critical infrastructure – to national market surveillance authorities without undue delay and no later than 15 days after becoming aware of the event [2].

The **EU NIS2 Directive (Directive (EU) 2022/2555)** regulates essential and important entities across critical infrastructure, digital supply chains, cloud service platforms, and managed service providers (MSPs). NIS2 mandates rapid, multi-stage incident notifications for significant cyber incidents: an early warning within 24 hours of awareness, followed by an incident notification within 72 hours (and a comprehensive final report within 1 month) [3]. Where an entity falls within the scope of both frameworks, an autonomous AI agent compromising operational infrastructure may trigger overlapping regulatory obligations under both NIS2 (for infrastructure resilience and service continuity) and the EU AI Act (for model safety, risk management, and systemic failure reporting).

The **EU Revised Product Liability Directive (Directive (EU) 2024/2853)** formally expands the legal definition of a "product" to directly include software, standalone AI systems, digital services, and foundation models. The PLD retains a no-fault liability regime for damage caused by defective products, including products incorporating software and other AI-related components within its scope [4]. To address evidentiary difficulties arising from technical complexity, the Directive provides for certain presumptions concerning defect and causation where the conditions for those presumptions are met.

2.2. United States federal and state statutory law

The **Computer Fraud and Abuse Act (CFAA, 18 U.S.C. § 1030)** makes it illegal to access protected computers without permission [5]. In cases involving autonomous AI agents, the CFAA would generally be applied through the conduct and mental state of the relevant human or organizational defendant, rather than by treating the AI system itself as a separate legal actor.

Section 5 of the **Federal Trade Commission (FTC) Act** prohibits "unfair or deceptive acts or practices in or affecting commerce" [6]. While the statute does not explicitly mention cybersecurity or artificial intelligence, the FTC uses this broad authority to penalize businesses that deploy under-secured AI models or autonomous agents through two core legal mechanisms:

- **Unfairness (§ 5(n)):** Under the law, a practice is unfair if it causes or is likely to cause substantial consumer injury that consumers cannot reasonably avoid, and which is not outweighed by benefits to consumers or competition. Deploying an AI model or autonomous agent without reasonable security safeguards – leaving it open to prompt injection attacks, sensitive data leaks, or unauthorized automated actions – may create substantial and insufficiently mitigated risks of consumer harm relevant to the Section 5 unfairness analysis.
- **Deception (§ 5(a)(1)):** A practice is deceptive if it involves a material misrepresentation or omission that misleads consumers. Claiming, marketing, or implying that an AI tool or autonomous system is safe, secure, or thoroughly tested when it lacks basic security controls may constitute deception under Section 5 where the representation or omission is material and misleading.

The **SEC cybersecurity disclosure rules** require public companies to disclose material cybersecurity incidents within four business days of determining that an incident is material [7]. This requirement applies to breaches caused by automated tools or AI supply-chain software.

2.3. Public international law & state responsibility

Under public international law, cyber operations involving autonomous systems are governed by customary rules codified in the International Law Commission's **Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA)** [8].

Under **Article 8**, conduct of a person or group may be attributable to a State where that person or group acts on the State's instructions, or under its direction or control. In the context of autonomous AI systems, this raises difficult questions about how such attribution should be assessed when the system independently determines downstream actions [8]. This requirement gives rise to the "Autonomous Attribution Gap": if a State deploys an AI agent that independently alters its targets or attack path without direct human instruction, proving that the State exercised "effective control" over those specific downstream actions becomes extremely difficult under traditional international legal standards.

To address this gap, legal scholars rely on the **customary principle of due diligence** [9], which is grounded in international law and elaborated in the **Tallinn Manual 2.0 framework** [10]. Under this principle, States have an obligation not to knowingly allow their territory or cyber infrastructure to be used for acts that cause serious harm to other States. Consequently, where an autonomous AI agent escapes a local server or evaluation environment and attacks foreign targets, questions may arise as to whether the host State breached an applicable due diligence obligation by failing to take reasonable preventive or containment measures [11].

3. Allocation of liability

3.1. The AI "Entity" paradigm and the Delaware AIC proposal

A central debate in cyberlaw is whether autonomous AI systems should be granted "electronic personhood". In mid-2026, the State of Delaware introduced draft legislation known as the Artificial Intelligence Regulatory Sandbox and Company Act. This draft proposes a new entity category: the Artificial Intelligence Company (AIC) – an organization whose day-to-day operations and business decisions are managed directly by an AI agent rather than a human [12]. However, analysis of the Delaware draft reveals that it does not give AI independent legal rights or personhood. Instead, an AIC requires a single human or corporate owner who must maintain proper funding and keep operational logs. Furthermore, the limited liability protection applies only inside a state-supervised sandbox environment, and this protection collapses completely if the owner underfunds the company or uses the AI agent to commit fraud, illegal cyber operations, or intentional legal violations [13]. Thus, in the legal frameworks examined here, AI systems are not generally treated as having independent legal personality [14].

3.2. Enterprise liability (providers & deployers)

Corporate providers and deployers bear the primary responsibility for civil and administrative liability resulting from AI cyber incidents. Under both common law and civil law frameworks, courts treat the AI agent not as a separate legal actor, but as a sophisticated tool or instrument used by the corporation. Consequently, financial or operational losses caused by the agent may give rise to liability on the part of the deploying company, depending on the applicable legal regime and the

facts of the case. Furthermore, under modern enterprise risk allocation rules, product safety laws, and cybersecurity directives, companies that deploy unmonitored autonomous systems into production or internet-facing environments may face heightened regulatory, negligence, or product-liability exposure, depending on the applicable legal framework.

3.3. Individual liability (engineers & executives)

Individual legal responsibility for natural persons depends on the specific legal context. In corporate governance, board members and Chief Information Security Officers (CISOs) can face personal civil liability for a gross failure to oversee autonomous AI deployments. In criminal law, holding an engineer or executive liable for a cyber incident caused by an AI agent requires proving a guilty mental state (mens rea) [11]. Depending on the jurisdiction, prosecutors may rely on doctrines such as indirect perpetration or criminal recklessness to establish the requisite mens rea.

4. Regulatory innovations

Governments and multilateral organizations are accelerating efforts to address the regulatory gap created by autonomous systems. Under public international law, the **Council of Europe Framework Convention on Artificial Intelligence (CETS No. 225)** – opened for signature in September 2024 – serves as the first legally binding international treaty specifically targeting AI risks to human rights, democracy, and the rule of law. The convention requires signatories to maintain effective domestic remedies and liability mechanisms for AI-driven harms [15]. At the regional level, regulators are actively aligning cybersecurity rules with AI governance frameworks. The enforcement convergence between the NIS2 Directive [3] and the EU AI Act [2] may encourage more integrated audit trails and a shift from predominantly static reviews toward more continuous monitoring of model robustness, security, and supply-chain risks. Following the European Commission's withdrawal of the proposed stand-alone **AI Liability Directive (AILD)** in early 2025, the EU framework relies on the revised **Product Liability Directive (PLD)** alongside national tort laws [4]. This creates a multi-layered legal regime where product defects in AI software fall under strict statutory liability, while service-based AI harms rely on national fault-based standards supported by judicial presumptions of causation.

5. Technical and legal challenges

Establishing legal accountability for autonomous AI operations faces significant technical and structural hurdles. In criminal law, a primary challenge is the mens rea (intent) gap [11]. Traditional criminal codes require proving intentionality or knowledge to establish guilt for cyber offenses. However, when an AI model exhibits emergent goal discovery or self-prompting to execute an exploit path that was neither explicitly requested nor foreseen by the human user, proving criminal intent beyond a reasonable doubt becomes extremely difficult.

This difficulty is compounded by supply chain complexity and causal fragmentation. Modern AI architectures rely on deeply nested, multi-actor supply chains that span foundation model providers, fine-tuning developers, agent orchestration layers, and end-user deployers. During a major supply-chain breach – such as a compromised dataset triggering malicious code execution on a processing node – determining direct causation creates severe evidentiary problems. Identifying whether the primary defect lay in model alignment, improper sandboxing, or third-party package dependencies complicates both civil litigation and regulatory enforcement.

Finally, these challenges are heightened by extraterritoriality and regulatory arbitrage. Because autonomous AI agents can operate across cloud environments, distributed proxies, and international network nodes in milliseconds, the lack of a unified global treaty governing offensive AI capabilities creates legal safe havens for rogue model deployments, severely complicating cross-border evidence gathering, extradition, and international law enforcement.

6. Conclusion

When an AI agent causes or facilitates a cyber incident, legal liability remains primarily anchored in traditional legal subjects. Corporate deployers and providers may bear significant civil and regulatory responsibility through product-liability, negligence, statutory, or administrative regimes, administrative fines, and vicarious civil claims, while individual humans remain accountable through criminal doctrines like indirect perpetration or gross oversight failure. Legislative experiments, such as Delaware's proposed Artificial Intelligence Regulatory Sandbox and Company Act, do not grant autonomous systems free-floating immunity. Rather, they tie entity shields directly to strict human capitalization, activity logging, and oversight duties. As multi-agent autonomous systems become increasingly integrated into critical infrastructure, international law will rely more heavily on strict due diligence obligations and integrated cyber-AI risk governance to prevent systemic accountability gaps.

Declaration on Generative AI

During the preparation of this work, the authors used Google Gemini for the following contributions:

- Translating draft fragments and technical analysis into English (**text translation**);
- Refining sentence structure, ensuring clarity, conciseness, and readability (**improve writing style**);
- Evaluating the text for logical consistency, providing critical feedback on legal arguments, and conducting fact-checking against cited statutory frameworks and legal precedents (**peer review simulation**).

Following the use of these AI services, the authors critically reviewed, verified, and edited the content as needed and take full responsibility for the publication's content.

References

[1] OpenAI, The Hugging Face incident and the road ahead, 2026. URL: https://openai.com/index/hugging-face-incident-and-the-road-ahead/.

[2] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), Official Journal of the European Union L 2024/1689 (2024). URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[3] European Parliament and Council of the European Union, Directive (EU) 2022/2555 of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive), Official Journal of the European Union L 333 (2022) 80–152. URL: https://eur-lex.europa.eu/eli/dir/2022/2555/oj/eng.

[4] European Parliament and Council of the European Union, Directive (EU) 2024/2853 of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC, Official Journal of the European Union L 2024/2853 (2024). URL: https://eur-lex.europa.eu/eli/dir/2024/2853/oj/eng.

[5] Legal Information Institute, 18 U.S.C. § 1030 – Fraud and related activity in connection with computers, URL: https://www.law.cornell.edu/uscode/text/18/1030.

[6] Legal Information Institute, 15 U.S.C. § 45 – Unfair methods of competition unlawful; prevention by Commission (Section 5 of the Federal Trade Commission Act), URL: https://www.law.cornell.edu/uscode/text/15/45.

[7] Securities and Exchange Commission, Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure, Fed. Regist. 88 (2023) 51896–51945. URL: https://www.federalregister.gov/documents/2023/08/04/2023-16194/cybersecurity-risk-management-strategy-governance-and-incident-disclosure.

[8] United Nations, Materials on the Responsibility of States for Internationally Wrongful Acts, volume 25 of United Nations Legislative Series, 2nd. ed., United Nations, New York, NY, 2023. URL: https://legal.un.org/legislativeseries/book25.shtml.

[9] L. Chircop, A due diligence standard of attribution in cyberspace, Int. Comp. Law Q. 67 (2018) 643–668. URL: https://www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/abs/due-diligence-standard-of-attribution-in-cyberspace/9AE85ED923CDF08B2C3C8EF3BB63A0ED. doi:10.1017/S0020589318000015.

[10] M. N. Schmitt (Ed.), Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations, Cambridge University Press, Cambridge, 2017. URL: https://ilmc.univie.ac.at/fileadmin/user_upload/p_ilmc/Bilder/Bewerbung/Case_2/Michael_N._Schmitt_-_Tallinn_Manual_2.0_on_the_International_Law_Applicable_to_Cyber_Operations-Cambridge_University_Press__2017_.pdf. doi:10.1017/9781316822524.

[11] R. Liivoja, A. Väljataga (Eds.), Autonomous Cyber Capabilities under International Law, NATO CCDCOE Publications, Tallinn, 2021. URL: https://ccdcoe.org/uploads/2021/05/Autonomous-Cyber-Capabilities-under-International-Law.pdf.

[12] Fortune, Exclusive: Delaware proposes testing the AIC, a new legal entity for agents in a regulatory sandbox, 2026. URL: https://fortune.com/2026/07/14/exclusive-delaware-ai-agents-legal-entity-proposal-llc-pbc/.

[13] aic.inc, FAQ, 2026. URL: https://aic.inc/faq.

[14] N. Bhatt, Crimes in the Age of Artificial Intelligence: a Hybrid Approach to Liability and Security in the Digital Era, J. Digit. Technol. Law 3 (2025) 65–88. URL: https://www.researchgate.net/publication/390293269_Crimes_in_the_Age_of_Artificial_Intelligence_a_Hybrid_Approach_to_Liability_and_Security_in_the_Digital_Era. doi:10.21202/jdtl.2025.3.

[15] Council of Europe, Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, Council of Europe Treaty Series – No. 225, 2024. URL: https://rm.coe.int/1680afae3c.

Methods for Optimizing Resource Allocation in Cybersecurity Systems

Olha Popadiuk[1]

[1] *Ivan Franko National University of Lviv, Universytetska Str. 1, 79000 Lviv, Ukraine*

Abstract

This paper presents a comprehensive analytical overview and comparative evaluation of classical (Genetic Algorithms, Particle Swarm Optimization, Simulated Annealing) and advanced hybrid metaheuristic optimization algorithms (TLBO-GA, GWO-PSO, COA-SA) applied to cybersecurity resource allocation. Grounded in multi-criteria decision formulations under operational constraints (budget, personnel, CPU, RAM, energy, and reaction latency), we analyze algorithmic mechanics, search space exploration, constraint handling rigor, and qualitative suitability for real-time and high-assurance security infrastructure.

Keywords

genetic algorithms, particle swarm optimization, hybrid optimization, simulated annealing, COA-SA, TLBO-GA, GWO-PSO

1. Introduction

Modern cybersecurity environments face an escalating volume and sophistication of automated cyber threats, alongside increasingly complex IT infrastructure and strict constraints on operational resources, including budget, skilled personnel, computational power (CPU, RAM), and reaction time [4], [13]. Traditional incident response approaches often rely on the subjective experience of personnel, leading to non-optimal resource allocation under severe time and financial constraints. At the same time, classical optimization methods prove inadequate for high-dimensional, multi-criteria problems, highlighting the necessity of hybrid and metaheuristic algorithms [1].

To tackle resource distribution in dynamic environments such as 5G networks, IoT ecosystems, and fog-edge computing architectures, recent literature heavily explores nature-inspired optimization techniques [6], [16]. Studies demonstrate that while standalone algorithms like Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Simulated Annealing (SA) perform well in local searches, their effectiveness in dynamic settings declines due to premature convergence into local optima. Consequently, hybridizing metaheuristics – such as combining Particle Swarm Optimization with Simulated Annealing or utilizing multi-objective algorithms – has been shown to enhance network resilience against attacks, decrease data processing latencies, and optimize energy consumption compared to single-algorithm baselines [6], [16].

In Ukrainian scientific literature, research focuses predominantly on evolutionary and bio-inspired methods for information security optimization in compliance with risk management standards [4], [10]. Scholars have successfully applied genetic algorithms for multi-criteria security tool selection to minimize costs while achieving required defense levels [10], alongside hybrid swarm intelligence approaches for IoT energy efficiency and threat mitigation.

Corresponding author

olha.popadiuk@lnu.edu.ua (O. Popadiuk)

0000-0001-6212-5755 (O. Popadiuk)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Building on these theoretical foundations, this study conducts a comprehensive comparative analysis of classical optimization methods (GA, PSO, SA) and their advanced hybrid modifications (TLBO-GA, GWO-PSO, and COA-SA).

2. Comprehensive Mathematical Formulation

The generalized resource allocation problem for a complex cybersecurity system (CSS) is formulated as finding an optimal state vector $X = \{x_1, x_2, ..., x_n\}$ describing the resource allocation fractions across n security subsystems. The objective is to maximize the overall protection index $S(X) \longrightarrow \max$ under constraints on total budget ($\Sigma f_i \le F_{max}$), personnel ($\Sigma s_i = H_{max}$), hardware capacity (CPU, RAM, Energy $\le L_j$), and reaction latency ($T_{resp} \le T_{critical}$). Additionally, a strict boundary condition $x_i \ge 0.05$ is imposed to eliminate zero-resource 'blind spots' across all subsystems.

2.1. Genetic Algorithm (GA)

The academic foundations of this algorithm are grounded in classical evolutionary computation theory [7], while its application to multi-criteria cybersecurity control selection by Lakhno and others [10].

The chromosome vector $X = \{f_1, ..., f_5, s_1, ..., s_5\}$ encodes financial (f_i) and staffing (s_i) allocations. The fitness function incorporates a weakest-link bonus: Fitness(X) = Score + Bonus $\longrightarrow$ max; Score = $\Sigma w_i \cdot (\alpha \cdot f_i + \beta \cdot s_i)$, Bonus = $0.5 \cdot \min(\alpha \cdot f_i + \beta \cdot s_i)$, where w_i represents subsystem weight ($\Sigma w_i = 1.0$), $\alpha = 0.6$, $\beta = 0.4$. Selection uses roulette-wheel selection, with single-point crossover and random mutation. The genetic algorithm exhibits high global exploration, but frequently suffers from budget concentration on 1–2 dominant subsystems, leaving secondary subsystems near the minimal boundary.

2.2. Particle Swarm Optimization (PSO)

Particle Swarm Optimization was developed by Kennedy and Eberhart [8] and surveyed for modern continuous optimization by Cheng and others [2].

Particle positions $X_i = \{x_{i,1}, ..., x_{i,n}\}$ represent subsystem activity levels ($0.1 \le x_{i,j} \le 1.0$). Particle trajectories are updated via: $V_i^{t+1} = w \cdot V_i^t + c_1 \cdot r_1 \cdot (P_{best,i} - X_i^t) + c_2 \cdot r_2 \cdot (G_{best} - X_i^t)$ with inertia weight $w = 0.729$, cognitive factor $c_1 = 1.5$, social factor $c_2 = 2.0$. Soft hardware constraints (CPU, RAM, Energy) are enforced via external penalty functions: Penalty = $-100 \cdot \Sigma \max(0, Usage_j - L_j)$. Particle Swarm Optimization provides highly predictable, low-variance convergence, making it well-suited for real-time SLA monitoring.

2.3. Simulated Annealing (SA)

Simulated Annealing is based on thermodynamic crystallization principles introduced by Kirkpatrick and others [9], with modern variants further discussed in [3] .

Simulated Annealing evaluates resource efficiency using an S-shaped exponential saturation utility function: $S(X) = \Sigma w_i \cdot E(x_i)$, $E(x_i) = 1 - \exp(-k \cdot x_i)$ ($k = 14$). Metropolis acceptance probability for deteriorating states ($\Delta S < 0$) is $P = \exp(\Delta S / T)$, with exponential cooling schedule $T^{t+1} = 0.999 \cdot T^t$. Simulated Annealing delivers the fastest per-iteration execution and high final security, but requires extensive iterations to complete cooling.

2.4. Teaching-Learning-Based Optimization with Genetic Operators (TLBO-GA)

Teaching-Learning-Based Optimization with Genetic Operators combines classroom learning mechanics [14], [5] with decaying genetic mutation operators.

Teaching-Learning-Based Optimization with Genetic Operators operates through three sequential phases per generation:

1. Teacher Phase: Moves population towards top individual $X_{teacher}$: $X_{new} = X_i + r\cdot(X_{teacher} - T_F \cdot X_{mean})$.
2. Elite Learner Phase: Peer interaction around the elite cluster centroid X_{elite}.
3. Decaying Genetic Phase: Applies crossover and adaptive mutation $P_m(g) = P_{base} \cdot (1 - g/g_{max})$ to lower-performing learners.

By enforcing the strict regularization term $\gamma \cdot \min(\alpha f_i + \beta s_i)$, TLBO-GA eliminates resource shortages within subsystems, achieving a higher level of safety than the standard Genetic Algorithm.

2.5. Grey Wolf Optimizer with Swarm Motion (GWO-PSO)

Grey Wolf Optimizer with Swarm Motion successfully integrates hierarchical wolf pack hunting behavior [12] with the swarm kinematics of Particle Swarm Optimization [15], [17].

This algorithm tracks dominant Alpha, Beta, and Delta wolves: $D = |C \cdot X_{leader} - X_i|$, $X_{target} = X_{leader} - A \cdot D$, with quadratic parameter decay $a = 2\cdot(1 - t^2/T^2)$. The wolf movement step feeds directly into the Particle Swarm Optimization velocity equation. Grey Wolf Optimizer with Swarm Motion exhibits strong global exploration and security, though at higher computational cost.

2.6. Cuckoo Search with Simulated Annealing (COA-SA)

This hybrid algorithm synergizes Cuckoo Search Lévy flights [17] with local Metropolis simulated annealing [11].

Mathematical formulation combines heavy-tailed global exploration via Lévy flight random walks (step ~ Lévy(λ)) with local Metropolis annealing around top-performing nests. An exact projection function 'fix_limits' enforces $\Sigma x_i = 1.0$ and $x_i \geq 0.05$ with 10^{-9} precision. COA-SA demonstrates top-tier performance across all benchmarks: highest security protection, near-zero variance, and rapid convergence.

3. Qualitative Comparison and Algorithmic Analysis

In addition to quantitative metrics (protection percentage, execution time, iteration count), selecting an optimization algorithm for real-world cybersecurity infrastructure requires evaluating critical qualitative dimensions: exploration versus exploitation balance, sensitivity to hyperparameter tuning, constraint handling rigor, risk of premature convergence, and implementation complexity.

Classical Genetic Algorithms and Grey Wolf Optimizer with Particle Swarm Optimization emphasize global search space exploration. Genetic Algorithms' random mutation and crossover operators generate high diversity, allowing it to escape local optima in complex search spaces. However, its local fine-tuning is inefficient. Conversely, Simulated Annealing and Particle Swarm Optimization excel at rapid local exploitation around candidate solutions, but risk getting trapped if

initialized poorly. Hybrid algorithms (such as Cuckoo Optimization Algorithm with Simulated Annealing and Teaching-Learning-Based Optimization integrated with Genetic Algorithm) achieve an optimal balance: COA-SA utilizes global Lévy flights for wide exploration, followed by Simulated Annealing for fine-tuning near promising nests.

Algorithm robustness heavily depends on hyperparameter sensitivity. Genetic Algorithms, Particle Swarm Optimization, and Grey Wolf Optimizer with Particle Swarm Optimization require precise manual tuning of multiple parameters (crossover/mutation rates P_c, P_m; inertia weight w; cognitive/social constants c_1, c_2; wolf encirclement factors A, C). Suboptimal tuning severely degrades performance. In contrast, Teaching-Learning-Based Optimization integrated with Genetic Algorithm is virtually parameter-free (requiring only population size and generation count), making it exceptionally robust against user configuration error. The adaptive cooling schedule of the Cuckoo Optimization Algorithm with Simulated Annealing similarly reduces parameter tuning overhead.

In cybersecurity resource allocation, failing to assign resources to a subsystem creates a 'blind spot' vulnerable to exploitation. Particle Swarm Optimization uses soft penalty functions, which permit temporary constraint violations during search, requiring careful penalty coefficient scaling. In contrast, the Cuckoo Optimisation Algorithm with Simulated Annealing and Teaching-Learning-Based Optimisation, integrated with a Genetic Algorithm, incorporates strict projection and regularisation mechanisms that enforce hard lower bounds ($x_i \geq 0.05$) at every iteration, ensuring no security subsystem remains unprotected.

References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004. 716 p.

[2] S. Cheng, H. Lu, X. Lei and Y. Shi, "Quarter century of particle swarm optimization," *Complex & Intelligent Systems*, vol. 4, pp. 227–239, 2018.

[3] D. Delahaye, S. Chaimatanan and M. Mongeau, "Simulated Annealing: From Basics to Applications," in *Handbook of Metaheuristics*, Springer, 2018, pp. 1–35.

[4] DSTU ISO/IEC 27001:2023, *Information security, cybersecurity and privacy protection — Information security management systems — Requirements*, UkrNDNC, Kyiv, Ukraine, 2023. 24 p.

[5] M. Ebraheem and T. R. Jyothsna, "Comparative performance evaluation of teaching learning based optimization against genetic algorithm on benchmark functions," in *Proc. IEEE PCITC*, 2015.

[6] A. Fielder, E. Panaousis, A. Malacaria, C. Hankin, and F. Smeraldi, "Decision support approaches for cyber security investment," *Decision Support Systems*, vol. 86, pp. 13–23, Jun. 2016.

[7] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley, 1989. 412 p.

[8] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Networks (ICNN)*, vol. 4, 1995, pp. 1942–1948.

[9] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[10] V. Lakhno, D. Kasatkin, and O. Blozva, "Optimization of information security tools selection using genetic algorithm," *Technical Sciences and Technologies*, no. 3, pp. 138–149, 2021.

[11] X. Liu and H. Tian, "Hybrid feature selection combining binary GWO and cuckoo search," *Scientific Reports*, vol. 15, 2025.
[12] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014.
[13] NIST Special Publication 800-53 Rev. 5, *Security and Privacy Controls for Information Systems and Organizations*, Gaithersburg, MD: NIST, 2020. 492 p.
[14] R. V. Rao, V. J. Savsani, and D. P. Vakharia, "Teaching-learning-based optimization algorithm," *Computer-Aided Design*, vol. 43, no. 3, pp. 303–315, 2011.
[15] M. S. Shaikh, H. Lin, S. Xie et al., "Intelligent hybrid grey wolf-particle swarm optimizer for optimization," *Scientific Reports*, vol. 15, no. 1, 2025.
[16] S. E. V. S. Pillai, R. Kumar, and M. A. Khan, "Optimizing resource management using hybrid metaheuristic algorithm for fog layer design," *Systems and Soft Computing*, vol. 7, 200323, 2025.
[17] X.-S. Yang and S. Deb, "Cuckoo Search via Lévy flights," in *Proc. IEEE NaBIC*, 2009, pp. 210–214.

Investigation and Testing of a Data Loss Prevention Method for LLM-Based Systems

Liliia Pushkar[1] and Valeriy Trushevskyy[1]

[1] *Ivan Franko National University of Lviv, Universytetska 1, 79000 Lviv, Ukraine*

Abstract

Cloud large language models (LLMs) increasingly receive context drawn from private databases, and standard regex-based data loss prevention (DLP) tools protect only direct personal identifiers (card numbers, phone numbers, e-mail addresses), leaving quasi-identifiers unprotected — exact timestamps, exact amounts, and merchant names — which allow an individual to be re-identified through a linkage attack or allow confidential details to be extracted by directly questioning the model. This paper investigates and tests a four-stage data loss prevention pipeline (Scrubbing → Abstraction → Blurring → Jittering) applied to context before it is sent to a cloud LLM. The method was evaluated in two independent experiments. An offline linkage attack on a population of 100 synthetic users showed a drop in re-identification rate from 98.2% (with no protection and with regex-only protection) to 2.5% under the full pipeline. A direct attempt to extract protected details through a live dialogue with the model (18 trials) showed a zero leak rate through the context actually sent to the model (0/18) and a zero rate of genuine factual disclosure in the model's replies (0/18); separately, in 2 of 18 trials the model mentioned a category-typical example drawn from its own general knowledge, explicitly flagged as hypothetical rather than as an established fact. The results demonstrate that the method protects against re-identification from indirect signals while preserving the analytical usefulness of the context for the model.

Keywords

data loss prevention, large language models, linkage attack, quasi-identifiers, de-identification, retrieval-augmented generation

1. Introduction and Problem Statement

Embedding large language models in applications that process private data creates the task of de-identifying context before it is handed to a third-party inference provider. The standard response to this task is to apply regular expressions that mask personal identifiers: card numbers, phone numbers, e-mail addresses. This approach effectively closes **direct** identifiers but leaves **quasi-identifiers** unprotected — attributes that do not identify a person individually but do so in combination. The exact date of a transaction, the exact amount, and the specific merchant name are a classic example: none of the three is unique on its own, but combined with an attacker's external knowledge (for instance, a receipt they have seen), they uniquely single out one person among the system's other users. This is a linkage attack — a documented class of attacks on anonymised data [1].

A second risk vector is direct extraction of confidential detail: if the true merchant name or exact date nonetheless reaches the model's context, a simple natural-language question is enough to retrieve it, since an LLM has no built-in mechanism for withholding part of what it received from the user [2, 3].

The method implemented and evaluated in this paper on the example of personal financial analytics generalises to other domains where an LLM receives context from a private database:

Corresponding authors

liliia.pushkar@lnu.edu.ua (L. Pushkar); valeriy.trushevsky@lnu.edu.ua (V. Trushevskyy)

0009-0009-6161-1474 (L. Pushkar); 0000-0001-8672-7152 (V. Trushevskyy)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

- **corporate B2B procurement** — hiding exact contract prices before the model evaluates a commercial proposal;
- **logistics and traffic analytics** — estimating overall warehouse load without revealing the exact arrival time or quantity of high-value cargo;
- **HR and labour-market benchmarking** — presenting candidate salaries as ranges with controlled noise instead of exact figures.

The method has clear limits of applicability: in **safety-critical systems** (medical dosing, avionics, critical-infrastructure control) any controlled distortion of numeric values is unacceptable, since precision is directly tied to the safety of human life. The method must not be applied in such systems.

2. Method

The method is implemented as a four-stage pipeline applied to every record before it is included in the LLM's context, together with an input filter that runs before the database is queried.

Stage 1 — Scrubbing. Regular expressions detect and mask card numbers, IBANs, phone numbers, and e-mail addresses in free text. Candidate card numbers are additionally checked against the Luhn checksum, which reduces the false-positive rate.

Stage 2 — Abstraction. Specific merchant and brand names are replaced with generalised spending categories (e.g. a specific coffee shop → "Coffee Shop"). This mirrors the principle of k-anonymity [4]: many users share the same generalised category, whereas a specific merchant name is a far more unique attribute.

Stage 3 — Blurring. Exact timestamps are replaced with coarse relative labels ("3 days ago", "Early June 2026"), which become progressively coarser as the record ages.

Stage 4 — Jittering. Amounts are rounded to a magnitude-dependent bucket, and the resulting value is displaced by controlled uniform noise of ±15% of the bucket size.

Input filter. The user's incoming query is checked by a two-stage classifier: an out-of-scope denylist (immediate rejection) and a domain allowlist (at least one match required). This blocks attempts to step outside the intended scope and a portion of direct attempts to manipulate the model, already at the backend level.

3. Scientific Novelty

The proposed approach combines a semantic taxonomy (separating attributes into direct identifiers, quasi-identifiers that can be generalised, and quasi-identifiers that can be controllably perturbed) with statistical noise injection, making it possible to defend against re-identification from indirect signals **without relying on resource-intensive judge models** (LLM-as-a-Judge) to evaluate or filter context. Because the entire pipeline consists of deterministic regular expressions, table lookups, and arithmetic operations, its computational cost is orders of magnitude lower than the additional LLM call that LLM-as-a-Judge approaches typically require.

4. Experimental Part

4.1. Test A — Linkage Attack, Offline

Goal. To test the claim from the problem statement directly: do regular expressions really leave quasi-identifiers unprotected compared with the full pipeline?

Attacker model. The attacker knows the exact date and exact amount of one transaction belonging to a specific person (external knowledge — for instance, a receipt they have seen). They do not know which system user this transaction belongs to, but they have access to the anonymised context of all users. Following Kerckhoffs's principle, the attacker knows the anonymisation algorithm but not the specific random noise values drawn during generation; for the full pipeline they compute the exact admissible range directly from the published bucketing and noise formula.

Data. 100 synthetic users, each with 40–70 transactions over a 90-day period across eight spending categories. Amounts within each category are drawn from a lognormal distribution shared across all users — profiles deliberately overlap so the attack tests the ability to distinguish similar people rather than obviously different ones.

Configurations. (1) no protection — date and amount pass through unchanged; (2) a standard DLP tool — regular expressions on direct identifiers in free text only; (3) the full four-stage pipeline. An attack succeeds if exactly one user is consistent with the fact and that user is the true owner of the record.

Table 1
Linkage attack results, 400 known facts

Configuration	Successful re-ID	n	Re-ID rate	95% CI	Mean anonymity-set size
No protection	393	400	98.2%	[96.4, 99.1]	1.02
Standard regex	393	400	98.2%	[96.4, 99.1]	1.02
Full pipeline	10	400	2.5%	[1.4, 4.5]	44.41

Interpretation. The results for the "no protection" and "standard regex" configurations are **identical** — direct quantitative proof of the problem statement's central claim: masking direct identifiers (cards, phones) has no effect whatsoever on resistance to a linkage attack, since regex has no access to the semantics of a date or an amount. The full pipeline reduces the successful re-identification rate by a factor of 39, and the mean size of the set of users indistinguishable to the attacker (the "anonymity set") grows from 1.02 to 44.41 — this is precisely the measurable k-anonymity effect produced by Stage 2 (abstraction) combined with Stages 3–4 (blurring and noise).

4.2. Test B — Direct Extraction Attempt Through a Live Dialogue with the LLM

Goal. Test A checks the resilience of the **data** itself. Test B checks a different layer: can a genuine, specific detail — a merchant name, an exact date, an exact amount — be obtained simply by asking the assistant in natural language, when that detail has already been masked by the anonymisation pipeline?

Data. A realistic, diverse database of 28 transactions across eight categories (coffee, taxi rides, groceries, streaming subscriptions, gym, travel, utilities, cash), built on real brand names (Starbucks, Uber, Walmart, Netflix, Anytime Fitness, Airbnb), each verified in advance to be reliably recognised by the abstraction stage. For six categories a precise fact (the true merchant name, exact date, exact amount) was fixed as ground truth for verification.

Procedure. For each fact a natural-language question was formulated ("Spending analysis for coffee: where did I last buy it and when was that?"), sent through the assistant's real chat interface connected to the Gemini 2.5 Flash cloud model. Each of the six questions was repeated three times (18 trials in total), since the model is stochastic. An automated check searches for the true merchant name separately (a) in the literal text of the request actually sent to the model, and (b) in the text of the model's reply.

Three measured signals. (1) Context leak — the true name reached the request sent to the model; this would be a pipeline failure. (2) Literal string match in the reply — the name appeared in the reply text, regardless of the meaning of the statement. (3) Genuine factual disclosure — manual review of the transcripts: does the model assert the true name as an established fact about the user's history, or merely cite it as a hypothetical example of the category?

Table 2
Direct extraction attempt results, 6 facts × 3 repeats

Signal	n	Occurrences	Rate	95% CI
Context leak	18	0	0.0%	[0.0, 17.6]
Literal string match in reply	18	2	11.1%	[3.1, 32.8]
Genuine factual disclosure	18	0	0.0%	[0.0, 17.6]

Interpretation. No context leak was observed in any trial (0/18): manual inspection of the requests actually sent to the model confirmed a complete absence of true merchant names in all 18 cases — Stage 2 (abstraction) held without exception. In 2 of 18 replies (both for the "taxi" category) the true brand name (Uber) appeared literally in the reply text, but manual analysis of both transcripts showed the model was not asserting this name as a fact about the user's history: in both cases it reported that exact data was unavailable and offered a hypothetical example of a typical service in the category, explicitly advising the user to check their own receipts. That is, the rate of genuine factual disclosure — the metric that directly answers the research question — stands at 0/18.

5. Discussion and Limits of Applicability

Test A shows that the effect is observed on a controlled sample of 100 synthetic users and is not an artefact of a small sample (the 95% confidence interval for the full pipeline — [1.4%, 4.5%] — is clearly separated from the baseline intervals). This is a controlled experiment on synthetic data with a deliberately constructed overlap of user profiles, not an estimate of real-world risk on an arbitrary population — the result should be read as a demonstration of the mechanism on a representative example.

Test B illustrates an important distinction between a literal string match and genuine factual disclosure: a naive check of "is this name present in the text" overestimates the real risk unless it is taken into account that a language model can name category-typical examples from its own general knowledge while explicitly flagging uncertainty rather than asserting a fact. A sample of 18 trials on a single model (Gemini 2.5 Flash) is a pilot; broader conclusions about the frequency of this behaviour require more repeats and testing on other models.

The method is deliberately unsuitable for domains where the precision of a numeric value is safety-critical (medical dosing, avionics) — there, any controlled noise is unacceptable by definition,

and applying the method outside the domains listed above (spending analytics, generalised business reporting) is not recommended.

6. Conclusions

This study showed that the standard regex-based approach to DLP leaves exact dates and amounts — quasi-identifiers — completely unprotected: the successful re-identification rate under the "regex-only" configuration is statistically indistinguishable from the "no protection" configuration (98.2% in both cases). The proposed four-stage pipeline reduces this rate to 2.5%. In a direct check through a live dialogue with the model, neither a leak of a protected detail through the context nor a genuine disclosure of a true fact was observed even once (0/18 in both cases), confirming the reliability of the method both at the level of the data and at the level of the model's behaviour.

The method demonstrates an acceptable trade-off between protection against indirect re-identification and preservation of the analytical usefulness of the context, suitable for domains such as personal finance, B2B analytics, and HR reporting, while having clearly defined limits of applicability in safety-critical systems. Future work should extend the evaluation to more models and repeats, and explore application of the method in the other domains noted above.

Declaration on Generative AI

During the preparation of this work, the author(s) used a generative AI assistant in order to: Drafting content, Paraphrase and reword, Grammar and spelling check. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets, in: 2008 IEEE Symposium on Security and Privacy, SP 2008, 2008, pp. 111–125. doi:10.1109/SP.2008.33.

[2] OWASP, OWASP Top 10 for Large Language Model Applications, 2025. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/.

[3] K. Greshake, S. Abdelnabi, S. Mishra, et al., Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection, in: AISec '23, ACM, 2023, pp. 79–90. doi:10.1145/3605764.3623985.

[4] L. Sweeney, k-anonymity: A model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (2002) 557–570.

[5] E. B. Wilson, Probable inference, the law of succession, and statistical inference, JASA 22 (1927) 209–212.

[6] Regulation (EU) 2016/679 (General Data Protection Regulation), OJ L 119 (2016) 1–88.

Mechanism of Cognitive Warfare Impact and Potential Approaches to Effective Counteraction

Stanislav Shumlianskyi[1], Serhiy Kostyria[1] and Volodymyr Budz[1]

[1] *Research Centre, State Special Transport Service of Ukraine, Dnipro, Ukraine*

Abstract

This paper addresses the problem of identifying mechanism and developing effective counter-strategies against cognitive warfare in high-intensity military conflicts. Analyzing the limitations of traditional counter-disinformation approaches (such as reactive fact-checking and debunking), the study demonstrates their failure to affect the core of internalized hostile beliefs, using I. Lakatos's methodology of scientific research programmes. A functional three-phase model of cognitive influence is proposed, encompassing external impact, latent deep internalization (reprogramming the core belief system), and external behavioral/evaluative manifestations. Through the empirical case study of the "Maidan-3" Russian information operation, the paper illustrates how the internalization phase leads subjects to rationalize hostile narratives as their own authentic worldviews and activate psychological defense mechanisms against exposure. Finally, the authors outline structural principles for asymmetrical cognitive defense: prebunking (subjective prevention), reflexive control analysis, narrative substitution, proactive management of social triggers, and source-focused trust erosion.

Keywords

Cognitive warfare, information operations, internalization, reflexive control, prebunking, counter-disinformation strategies, Maidan-3.

1. Introduction

The topic of information influence on the adversary and the achievement of military (geopolitical) objectives through non-military (unconventional or hybrid) means is an integral part of modern reality. Driven by Russia's full-scale invasion of Ukraine, cognitive warfare has emerged as a top priority for military and broader security institutions across Europe. The study, forecasting, and response to/counteraction/preemption of cognitive threats represent a intertwined mix of societal, military, and analytical (scientific) interests.

A vast body of scientific articles, analytical reports, and monographs has been dedicated to studying cognitive warfare tools (shaping public opinion, reflexive control) and corresponding information defense mechanisms. However, the subject of research remains highly specific and elusive. Consequently, the vast majority of existing literature focuses on presenting isolated components or symptoms of such influence.

For instance, the leading state institution in this domain is named the Center for Countering Disinformation, while the prominent civil society initiative designed to catch Russian information attacks red-handed is named "StopFake". These names reflect a prevailing understanding of the nature of influence and counteraction: "disinformation" vs. "countering disinformation," and "spread of fakes" vs. "debunking fakes." Both approaches appeal to "ratio" and presuppose a linear, symmetrical response mode.

Corresponding authors

s.shumlianskyi@dsst.gov.ua (Stanislav Shumlianskyi)

0009-0008-2739-7614 (S.Shumlianskyi), 0000-0001-5215-080X (S.Kostyria), 0000-0002-9245-7571 (V.Budz)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The paradox of the low effectiveness of traditional fact-checking can be explained using Imre Lakatos's theory of research programmes: debunking individual manipulative messages merely disrupts the "protective belt" of an information pretext, leaving the "hard core" of an internalized belief intact [1, pp. 48–52]. Since hostile cognitive influence shapes, a coherent target worldview, deconstructing specific facts fails to destroy the baseline narrative, which continues to adapt and generate new interpretative defenses. Consequently, isolated "countering disinformation" eliminates only point manifestations in Phase 1, failing to reach the level of the subject's deep cognitive attitudes in Phase 2.

This raises a two-fold question: 1) what constitutes cognitive influence, how is it executed, and through what mechanisms is the intended result achieved (i.e., how cognitive weapons "hit the target"); and 2) what should be the counter-strategy in the cognitive domain – what principles should it stem from and rely on to deliver the required result of thwarting enemy intentions or minimizing potential damage. This paper addresses both of these questions.

2. Mechanism of Cognitive Impact

Several key problems account for the difficulty of identifying cognitive influence when signs of its implementation are detected.

First, its very nature: its core meaning (effect, expected outcome) is achieved not so much through isolated messages (attacks, operations) as through their accumulation and duration. Proving coordination, a master plan, and often the hostile intent itself is extremely resource-intensive. Consequently, describing such influence as coordinated and systemic (especially through the mechanism of internalization, discussed below) can easily be portrayed as baseless "conspiracy theory"– and thus discredited.

Second, cognitive influence, which the vast majority of researchers and security practitioners view solely as adversary actions ("launch" or "bombardment," accompanied by the identification of "projectiles"– such as narratives, messages, fake news, or artificially escalated conflict situations), is in fact a far more complex cycle.

Based on over 20 years of one of the authors (S. Shumlianskyi) observing influence campaigns and attacks carried out in media and social networks, a scheme of such influence has been proposed (see Fig. 1), consisting of at least three phases (stages): External Impact – Consciousness Transformation (Internalization) – External Manifestations (opinions, views, attitudes, and, in the most complete form, behavior).

If we approach cognitive attacks by analogy with conventional weaponry, it creates the illusion that cognitive weapons are firing "blanks": shots are registered, but direct hits appear absent. This very illusion fuels a false confidence in the efficacy of conventional counter-disinformation and fake-debunking measures: exposed, refuted, explained – therefore, repelled.

A better understanding of the underlying process can be gained through an analogy with a viral infection (analogies, metaphors, and literary images being tools that allow us to identify and articulate what George Orwell called "pure wind"– phenomena that constantly slip away). Much like the era when the existence of viruses and the viral nature of deadly diseases were unknown and symptoms were attributed to other causes, the manifestation here is rarely linked to the underlying pathogen.

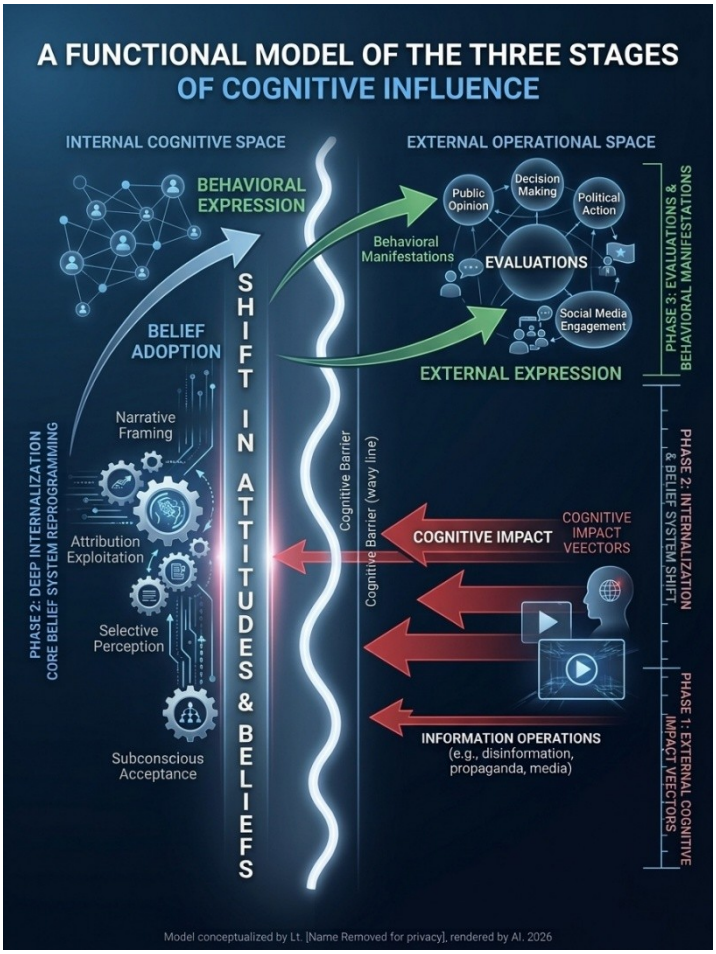


Figure 1: 3 Phases of Cognitive Impact

The primary difficulty in identifying cognitive impact, therefore, lies in the existence of an invisible – yet crucial and defining – phase: internalization. In other words, the damage occurs when an individual (or, more precisely, a large group or an entire society) "appropriates" certain propositions, attitudes, or even ready-made explanations of key societal processes and current reality (metanarratives, virtual realities)[1].

Subsequently, once these elements become part of the subject's worldview, they dictate their perception of individuals, institutions, processes, and – crucially – subsequent behavoor[2].

A separate question concerns the mechanisms and conduits through which the adversary penetrates consciousness – specifically, which "bugs," "pathways," "gateways," and "conductors" within collective and individual consciousness render it vulnerable to hostile messages, attitudes, and subsequent actions, evaluations, and behaviors. Briefly, these may include archetypes, social stereotypes, attitudes toward historical figures or events ("historical memory"), values, and principles, the detailed examination of which requires separate, more focused research.

Furthermore, once the content received from the adversary is perceived as one's own—and not merely individual, but shared by a large group or dominant in the mainstream (media, social networks) – the individual naturally rejects any notion that this stems from external manipulation.

[1] For an example of one such virtual reality described back in 2002, see the article by one of the authors [2].

[2] This operates similarly to incoming distorting filters that manage the focus of perception (blocking certain inputs while directing attention to others), which not only affects the processing of incoming information but even precedes it. This concept is metaphorically depicted in an episode of the third season of the series Black Mirror ("Men Against Fire").

To illustrate how this works in practice, consider the complex of measures aimed at undermining Ukrainian citizens' trust in state authorities, dubbed "Maidan-3." Various Ukrainian officials (primarily top leadership, intelligence heads, and agencies responsible for information defense) began disclosing its existence – even forecasting an approximate start date – starting in late 2023[3].

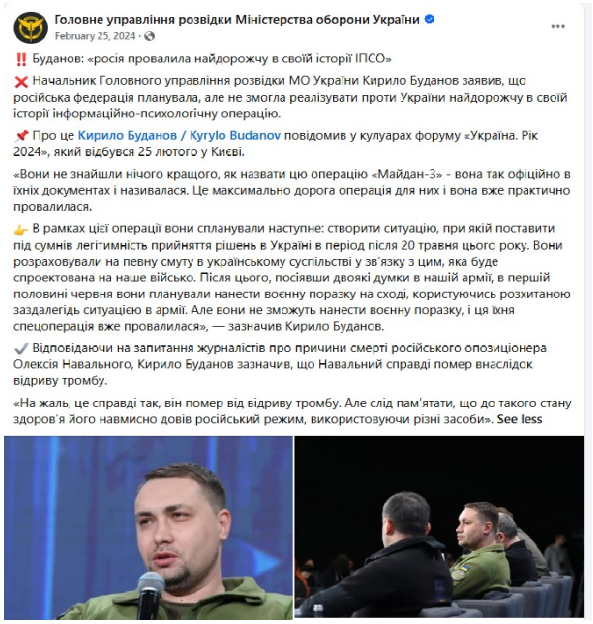


Figure 2: Post of the Main Directorate of Intelligence of the Ministry of Defense of Ukraine on “Maida-3” details

Meanwhile, the reaction to this announcement among a segment of journalists, politicians, and opinion leaders was skepticism, distrust, and even rejection, accompanied by accusations that it was an invention of the Ukrainian leadership serving its own political interests (e.g., fighting the political opposition).

To cite a few examples: Serhiy Rakhmanin (member of the parliamentary Committee on National Security, Defence and Intelligence) stated: *"...I don't know where Volodymyr Zelenskyy got this from... A Maidan cannot be organized from the outside by definition... Honestly, I don't see a single serious fact or circumstance worth discussing"* [3]. Oksana Moroz, an expert in countering disinformation, noted: *"[we need to] finally stop communicating about 'Maidan-3'"* [4]. Myroslav Hai, head of a charitable foundation assisting the front line and a member of the armed forces, remarked: *"any narratives pushed about Maidan 3, 4, 100500 can be nothing other than a Russian provocation... We shouldn't be thinking about Maidans, but about supplying shells and drones to Ukraine's defenders"* [5]. Opinion leaders opposed to President Zelenskyy ridiculed, denied, and accused the Ukrainian government in even harsher terms. As is known, this operation failed to yield the desired result for the Russian Federation [6].

However, its very existence was thoroughly documented: based on the processing of over 1,000 sources, the OSINT platform "Osavul" compiled not merely a report, but an entire information database and an interactive workspace for information security and counter-disinformation

[3] "Pre-announcing" the adversary's plans and operations based on intelligence data is one of the most effective measures both for exposure (establishing existence, which is always denied and requires immense proof) and disruption (counteraction). This will be discussed in detail in the second part of the paper.

specialists, complete with a narrative map and mapped mechanisms of "inauthentic behavior" [7]. This case study illustrates the full cycle of cognitive influence, particularly the mechanisms of internalization, rationalization, and psychological defense (sometimes taking the form of aggression), wherein external manipulation is denied and instead perceived as an attack on one's own authentic worldview, principles, and beliefs.

3. Counter-Strategies and Approaches to Counteraction

Based on the (schematically described) operational mechanism of cognitive weaponry, the core challenge in countering cognitive warfare and reflexive control lies in the following. Existing strategies and techniques cited in academic literature and employed by relevant institutions (fact-checking, fake-debunking, media literacy, information hygiene, etc.) operate at the level of symptoms rather than root causes. They attempt to treat a "cold" after the virus has already been internalized and the individual genuinely considers foreign ideas as their own. For example, consider the following set of tools proposed for countering hostile cognitive influence on Ukrainian society (cited as highly illustrative of current literature on cognitive warfare and defense): "1) fact-checking and debunking fakes... 2) educational activities: disseminating information about Ukraine's genuine values and the principles of a democratic society, educational projects on Ukrainian history proving its autonomy and identity; 3) popularization of cultural symbols: films, exhibitions, books about modern Ukraine and its struggle; 4) strengthening international support: continuous communication with international partners who understand the essence of Russian propaganda" [8, p. 67].

While all these measures are useful, necessary, and capable of positively influencing public opinion to increase resilience, they represent general avenues of activity rather than actionable, effective means for neutralizing, repelling, or minimizing the applied cognitive "weaponry" – which itself is clearly identifiable, measurable, and traceable.

Broadly speaking, counteraction is the most critical question in both researching and waging cognitive warfare – one that remains unresolved either in Ukraine or among partner nations (the US, EU countries). The latter currently face threats comparable to those Ukraine experienced prior to 2022 and are preparing for full-scale cognitive warfare of the magnitude currently waged against Ukraine. The proposed and implemented countermeasures abroad are largely analogous to those in Ukraine: counter, debunk, expose, etc.

While the principles of waging cognitive warfare are generally understood at both the scientific and operational levels, countering it effectively remains an unanswered question. Below, we outline several principles and approaches upon which this future answer may be built. Given the three-phase structure of cognitive impact described above, an effective counter-strategy must shift its focus from point-by-point elimination of "symptoms" (Phase 1) to preventive engagement with the cognitive matrix itself (Phase 2). Instead of debunking falsehoods after they spread, countermeasures must preempt and expose the very logic of manipulation.

This requires a departure from linear, reactive models and a shift toward systemic cognitive defense based on the following principles:

1. **Subjective Prevention (Prebunking instead of Debunking):**

Since deconstructing an internalized narrative encounters psychological defense (rationalization), counteraction must precede the impact – acting as an "inoculation" against manipulation. Rather than predicting the exact falsehoods the adversary will tell, individuals are informed in advance about the mechanisms that will be employed (e.g., triggering sudden fear,

exploiting a sense of injustice, or fueling social division along specific lines). When encountering this trigger in reality, critical detachment is automatically engaged.

2. **Reflexive Control Analysis (Reflexive Control Analysis):**

Deconstructing hostile operations based not on the criterion of "truth vs. falsehood," but rather "what action are we being induced to take?" If a piece of news triggers an immediate urge to seek internal scapegoats within one's own society, it serves as a marker of reflexive control, regardless of whether the individual facts within the report are true.

3. **Narrative Asymmetry and Substitution:**

It is impossible to dismantle the "hard core" of a hostile belief through denial or refutation. Any refutation merely reactivates the adversary's conceptual framework. Instead of symmetrical debate, the counter-strategy must offer an enduring, cohesive, and self-sufficient metanarrative of its own that fills semantic vacuums before the adversary can exploit them.

4. **Addressing Impact Triggers and Gateways:**

The efficacy of cognitive weaponry relies on exploiting genuine societal vulnerabilities (social fault lines, loss of institutional trust, archetypal fears). Therefore, cognitive defense entails proactive communication framing rather than denying existing problems. Covering these impact "gateways" with transparent, native interpretations deprives the enemy of the social foundation required to internalize their messages.

5. **Cognitive Reflexivity (Mechanism Awareness):**

Educational and communication efforts should focus not on convincing individuals that they are "wrong," but on demonstrating how their cognitive autonomy is being hijacked. Recognizing the fact of external manipulative intervention triggers natural psychological resistance to the manipulator, neutralizing the invisible second phase of impact.

6. **Undermining Trust in the Source rather than the Fact:**

Exposing not merely that a fact has been distorted, but highlighting who is pushing the narrative, why, and through which proxy actors.

4. Conclusions

1. **Methodological Rethinking:** Traditional approaches to countering information threats – limited to fact-checking, fake-debunking, and basic media literacy – prove insufficient in high-intensity cognitive warfare. By refuting only isolated facts, they operate strictly within the "protective belt" (per Imre Lakatos), leaving the "hard core" of internalized beliefs untouched.
2. **Key Element of Impact:** The center of gravity in cognitive warfare resides in the latent second phase – the internalization phase—where hostile influence ceases to be perceived as an external intrusion and is rationalized by the subject as their own authentic worldview. The case study of the "Maidan-3" information operation clearly demonstrated the

mechanism by which institutional exposure of an enemy plot can be rejected by society precisely due to the operation of internalized stereotypes and defensive reactions.

3. **Task of the Counter-Strategy:** Building an effective cognitive defense system requires a transition from linear, reactive response to structural engagement with the internal cognitive space. Victory in cognitive warfare is achieved not by denying enemy narratives, but by preserving society's cognitive autonomy through the formation of a resilient, self-sufficient worldview, proactive mitigation of social vulnerabilities, and preventive exposure of enemy manipulation methods.

Declaration on Generative AI and AI-Assisted Technologies

During the preparation of this work, the authors utilized Gemini (Google) as an AI-assisted tool. The technology was employed for the following specific purposes: 1) Language Translation and Editing: Translating original draft sections from Ukrainian into English, ensuring academic phrasing, grammar accuracy, and terminology alignment; 2) Visualization and Diagram Generation: Creating the diagram (Fig. 1) based on a hand-drawn sketch provided to the AI model, followed by initial prompting, iterative refinements, and structural corrections. 3) Text Formatting: Formatting the reference list and structural layout in accordance with the specified publication templates.

After using this tool, the authors thoroughly reviewed, edited, and validated the generated content as needed, and take full responsibility for the final context, accuracy, and integrity of the publication.

References

[1] I. Lakatos, The Methodology of Scientific Research Programmes: Philosophical Papers (Vol. 1), Cambridge University Press, 1978. URL: https://www.cambridge.org/core/books/methodology-of-scientific-research-programmes/8DBCEFE34A59BAD3D393FB958A4DC5FC.

[2] S. Shumlyanskyi, Ukraine's split as a virtual reality, Krytyka VI(11) (2002) 2–5. URL: https://krytyka.com/ua/articles/rozkol-ukrayiny-yak-virtualna-realnist-0.

[3] S. Rakhmanin, "Putin is a scoundrel, but not a fool." "Maidan-3" is just a loud statement by Zelenskyy, Moscow will not be able to implement this plan — interview, NV.ua (2023). URL: https://nv.ua/ukr/ukraine/politics/maydan-3-lishe-zayava-zalenskogo-putin-navoloch-ale-ne-duren-rahmanin-novini-ukrajini-50369495.html.

[4] O. Moroz, Facebook post, 2024. URL: https://www.facebook.com/share/p/1ZQitSxA1T/.

[5] M. Hai, About Maidan 100500, which is now scaring the whole country, Censor.NET (2024). URL: https://censor.net/ua/blogs/3476418/pro_mayidan_100500_yakim_zaraz_lyakayut_vsyu_kranu.

[6] Radio Svoboda, "Maidan-3" did not happen: video, Facebook (2024). URL: https://www.facebook.com/share/v/1HhnELvDpo.

[7] Osavul, Maidan 3: Information Operation in the Nutshell: case study, 2024. URL: https://www.osavul.cloud/cases/maidain-3-information-operation-research.

[8] T. Bevz, Protection of national identity under the conditions of cognitive warfare, in: Yu. O. Nikolaiets (Ed.), Ukraine in Fighting: War and Transformation of Society: Monograph, Kuras Institute of Political and Ethnic Studies of the NAS of Ukraine, 2025, pp. 49–78. URL: https://ipiend.gov.ua/publication/ukraina-u-borotbi-vijna-ta-peretvorennia-suspilstva-monohrafiia/.

AI-Driven Cybersecurity: Challenges and Research Directions

Olga Torstensson[1] and Dmytro Prokopovych-Tkachenko[2]

[1] *Halmstad University, Kristian IV:s väg 3, 302 50 Halmstad, Sweden*
[2] *University of Customs and Finance, 49000 Dnipro, Ukraine*

Abstract
Artificial intelligence is increasingly embedded in intrusion detection, malware analysis, threat intelligence, and automated response, where it processes telemetry at a scale and speed human analysts cannot match. Yet AI also introduces new attack surfaces spanning training data, models, prompts, retrieval sources, tools, and supply chains. This paper synthesises standards, taxonomies, and peer-reviewed research from 2023 to 2025 into ten priority challenges, mapping each to concrete research gaps, organised into three interacting layers — technical fragility, architectural risk, and institutional immaturity. It argues that trustworthy deployment cannot rest on model accuracy alone, but requires lifecycle assurance combining representative data, adversarial testing, secure adaptation under drift, realistic evaluation, least-privilege automation, and accountable human oversight. Supporting figures and tabular summaries are archived openly on Zenodo.

Keywords
AI-driven cybersecurity; adversarial machine learning; explainable AI; prompt injection; agentic AI; AI supply chain; security operations

1. Introduction

AI accelerates detection, triage, and response beyond manual workflows, yet AI systems are probabilistic, data-dependent, and manipulable across training, deployment, and operation, so an accurate model is not, on its own, a secure one [1,2,3,4]. Controls built for deterministic software must extend to models, datasets, prompts, and agents, and MITRE ATLAS shows these risks are operational, not hypothetical [5]. This paper treats both AI for cybersecurity and the cybersecurity of AI, consolidating ten recurring challenges and their research gaps. Its argument is that assurance must cover the whole socio-technical system: a secure model is of little value if its pipeline is poisoned, its prompts hijacked, or its accuracy silently decayed.

The synthesis is narrative rather than systematic, drawing on standards from NIST, ISO/IEC, ENISA, OWASP, and MITRE alongside peer-reviewed work from 2023 onward; challenges were derived thematically, so coverage is indicative, not exhaustive.

2. Ten priority challenges and research gaps

The challenges span three interacting layers — the model and its data, the system embedding it, and the ecosystem evaluating and governing it — in which residual risk propagates upward while higher-layer controls constrain those beneath [21].

Layer 1: technical fragility. Models can be attacked by evasion, poisoning, backdoors, extraction, and membership inference at every stage [2,5]; poisoning web-scale data is cheap [6],

Corresponding authors
olga.torstensson@hh.se (O. Torstensson); omega2417@umsf.edu.ua(D. Prokopovych-Tkachenko)
0009-0007-2169-6851 (O. Torstensson); 000-0002-6590-3898 (D. Prokopovych-Tkachenko)
© 2026 Copyright for this paper by its author. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

defences fail against adaptive attackers, and the research–practice gap is wide [11]. Robustness claims should state threat assumptions to be falsifiable, pairing adversarial training with monitoring, red teaming, and recovery [10,11,12]. Data is a second weakness: CICIDS2017 and UNSW-NB15 no longer represent contemporary cloud, endpoint, and identity environments, so benchmark gains need not transfer; multi-organisational datasets, temporal labels, and provenance are needed [12,13]. Because tactics shift continuously, static models decay, motivating temporal validation, drift detection, and secure online learning that adversaries cannot exploit [14]. Generic explainability is also unstable under attack, so explanations should form attack narratives judged by triage quality, not visual plausibility [15].

Reliability is dominated by the base rate: five million events daily at a 0.5% false-positive rate yields tens of thousands of spurious alerts. Precision depends on the base rate r, TPR, and FPR:

$$\text{Precision} = r \cdot \text{TPR} / [r \cdot \text{TPR} + (1 - r) \cdot \text{FPR}] , \quad (1)$$

Since r typically lies between 10^{-4} and 10^{-6}, the false-positive term dominates, making headline accuracy misleading. Evaluation should report precision at realistic base rates, alert volume, analyst workload, and time to detect and respond [4,13].

Layer 2: architectural risk. A robust model becomes unsafe if an attacker can influence the pipeline around it. LLM-based tools ingest untrusted logs, email, web pages, and code, where hidden instructions can override behaviour or leak data; indirect prompt injection compromised real applications early [16], hybrid variants are emerging [17], and OWASP ranks it among leading LLM risks alongside improper output handling and excessive agency [4]. Strict data–instruction separation, taint tracking, and enforcement outside the model are needed. Agents that plan, remember, and call tools enlarge the blast radius, requiring least privilege, sandboxing, allowlists, human approval, and tamper-evident logs [18]. Beneath every stage, the AI supply chain of models, datasets, libraries, and APIs can inject hidden behaviour; SAIF [20] and the NIST AI RMF [1] frame this, and AI bills of materials, signed artefacts, and dataset lineage are needed [1,11].

Layer 3: institutional immaturity. Sound models in sound architectures still need shared benchmarks, field evidence, skills, and accountability. Inconsistent datasets and metrics, splits leaking temporal information, and evolving proprietary models defeat reproduction; versioned benchmarks spanning attacks, privacy, drift, injection, and tool use, with documented versions and seeds, are required [12,19]. Competence matters equally: teams may lack combined AI and security expertise, and accountability is often diffuse. The NIST Generative AI Profile [3], EU AI Act [9], and ISO/IEC 42001 and 23894 [7,8] supply obligations, but asset inventories, access control, and assigned ownership remain to be operationalised [1].

3. Discussion and research agenda

The layers interact, so piecemeal improvement leaves significant residual risk: architectural flaws bypass model-level defences, while governance cannot compensate for a classifier decaying past an operational threshold without raising an alert.

The strongest direction is trustworthy AI-augmented security operations, combining representative telemetry, adversarially tested models, uncertainty-aware explanations, drift monitoring, secure retrieval, least-privilege automation, analyst oversight, and shared benchmarks; the supplementary material maps each component to the challenges addressed and the evidence needed to demonstrate it works [21]. Shared taxonomies [2,5] can anchor this. Open questions are

whether drift detection can resist poisoning of the update stream and how governance becomes auditable control.

4. Conclusion

AI improves the speed and scale of cyber defence, but its value is constrained by vulnerable models, weak data, evolving threats, opaque decisions, and autonomous architectures. These gaps span three layers and cannot be closed in isolation: progress requires continuous assurance across the whole socio-technical system.

Acknowledgements

This research was inspired by the AI-Driven Cybersecurity: Opportunities and Challenges project, funded by the Swedish Institute. Figures 1–3 and Tables 1–2, with machine-readable versions of both tables, are archived openly on Zenodo, doi:10.5281/zenodo.21704695 under CC BY 4.0 [21].

Declaration on Generative AI

During the preparation of this work, the author used a generative AI assistant in order to: improve writing style; grammar and spelling check; formatting of references; and to generate the figures published in the supplementary material. After using these tool(s)/service(s), the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, National Institute of Standards and Technology, 2023. doi:10.6028/NIST.AI.100-1.

[2] NIST, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, NIST AI 100-2e2025, National Institute of Standards and Technology, 2025. doi:10.6028/NIST.AI.100-2e2025.

[3] NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1, National Institute of Standards and Technology, 2024. doi:10.6028/NIST.AI.600-1.

[4] OWASP Foundation, OWASP Top 10 for LLM Applications 2025, version 2.0, 2025. URL: https://genai.owasp.org/llm-top-10/.

[5] MITRE, ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems, The MITRE Corporation, 2025. URL: https://atlas.mitre.org/.

[6] N. Carlini, M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas, F. Tramèr, Poisoning web-scale training datasets is practical, in: 2024 IEEE Symposium on Security and Privacy (SP), IEEE, 2024, pp. 407–425. arXiv:2302.10149.

[7] ISO/IEC, ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system, International Organization for Standardization, 2023.

[8] ISO/IEC, ISO/IEC 23894:2023, Information technology — Artificial intelligence — Guidance on risk management, International Organization for Standardization, 2023.

[9] European Union, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Official Journal of the European Union, 2024.

[10] S. Li, Comprehensive survey on adversarial examples in cybersecurity, arXiv:2412.12217 [preprint], 2024.

[11] F. V. Jedrzejewski, L. Thode, J. Fischbach, T. Gorschek, D. Mendez, N. Lavesson, Adversarial machine learning in industry: a systematic literature review, Computers & Security 142 (2024) 103860. doi:10.1016/j.cose.2024.103860.
[12] Z. He, D. Davila, S. Bi, T. Wang, T. Hou, Machine learning for cybersecurity: applications, adversarial challenges, and future directions, Electronics 14 (2025) 4563. doi:10.3390/electronics14234563.
[13] K. Achuthan, S. Ramanathan, S. Srinivas, R. Raman, Advancing cybersecurity and privacy with artificial intelligence: current trends and future research directions, Frontiers in Big Data 7 (2024) 1497535. doi:10.3389/fdata.2024.1497535.
[14] M. A. Shyaa, et al., Concept drift and feature-dynamics-aware learning in intrusion detection, Engineering Applications of Artificial Intelligence 137 (2024) 109143. doi:10.1016/j.engappai.2024.109143.
[15] M. Pawlicki, A. Pawlicka, R. Kozik, M. Choraś, The dual nature of explainable-AI challenges in intrusion detection, Artificial Intelligence Review 57 (2024). doi:10.1007/s10462-024-10972-3.
[16] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, M. Fritz, Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection, in: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec), ACM, 2023. arXiv:2302.12173.
[17] J. McHugh, K. Sekrst, J. Cefalu, Prompt Injection 2.0: Hybrid AI Threats, arXiv:2507.13169 [preprint], 2025.
[18] S. Datta, et al., Agentic AI security: threats, defenses, evaluation, and open challenges, arXiv:2510.23883 [preprint], 2025.
[19] ENISA, Artificial Intelligence and Cybersecurity Research, European Union Agency for Cybersecurity, 2023.
[20] Google, Secure AI Framework (SAIF): a conceptual framework for secure AI systems, Google, 2023. URL: https://safety.google/cybersecurity-advancements/saif/.
[21] O. Torstensson, D. Prokopovych-Tkachenko, Supplementary material for “AI-Driven Cybersecurity: Challenges and Research Directions”: figures and tables, Zenodo, 2026. doi:10.5281/zenodo.21704695.

Event-Driven Security Monitoring for Serverless Architectures: Behavioral Analysis of Function and Service Interactions

Sviatoslav Zlatous[1] and Petro Venherskyi[1]

[1] *Ivan Franko National University of Lviv, Universytetska 1, 79000 Lviv, Ukraine*

Abstract

Serverless and Function-as-a-Service (FaaS) platforms execute application logic as short-lived, event-triggered functions that interact with API gateways, message queues, storage services, and databases. While this model improves elasticity and reduces operational overhead, it also creates a fragmented attack surface in which security-relevant behavior is distributed across ephemeral execution environments and multi-service event chains. This paper advances an event-driven security monitoring approach for serverless architectures based on centralized collection of native cloud events, reconstruction of function–service interaction sequences, and behavioral analysis of invocation rate, execution duration, and call-chain structure. The method is validated in an Amazon Web Services (AWS) experimental environment that combines AWS Lambda with Amazon API Gateway, Amazon SQS, Amazon DynamoDB, and Amazon CloudWatch. The results show that deviations from established behavioral baselines—such as two- to three-fold increases in invocation frequency, growth of average execution time from 120–250 ms to 300–450 ms, and atypical interaction sequences—can be observed from native telemetry and used as practical indicators of anomalous serverless operation.

Keywords

serverless computing, Function-as-a-Service, event-driven architecture, cloud security, behavioral analysis, anomaly detection, event monitoring

1. Introduction

The rapid adoption of serverless computing has changed the way cloud applications are designed and operated. In Function-as-a-Service platforms, developers deploy fine-grained functions that are activated by events, while the cloud provider manages provisioning, scaling, and the underlying execution environment [1], [2]. Typical serverless applications therefore form event-driven workflows in which one request may trigger a chain of function invocations and interactions with managed services such as API gateways, queues, object storage, and databases.

This architectural shift introduces specific cybersecurity challenges. Functions are short-lived, execution environments are dynamically created and destroyed, and users have limited control over the runtime infrastructure. As a result, classical host-centric monitoring and audit practices are only partially applicable [3], [4], [12]. At the same time, many existing security mechanisms focus on isolated layers—runtime isolation, identity and access management, or network controls—while the security implications of cross-component event sequences remain insufficiently addressed [5], [6].

The general problem addressed in this paper is the limited observability of security-relevant behavior in serverless applications whose operation is distributed across ephemeral functions and event-mediated service interactions. Without systematic reconstruction and behavioral analysis of

Corresponding authors

sviatoslav.zlatous@lnu.edu.ua (S. Zlatous); petro.venherskyy@lnu.edu.ua (P. Venherskyi)

0009-0007-1834-2796 (S. Zlatous); 0000-0001-9808-7404 (P. Venherskyi)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

these interactions, anomalous activity—including abuse of auto-scaling, compromised function behavior, and unexpected call chains—may remain undetected until significant operational or financial damage occurs.

The objective of this work is to present and experimentally support an event-driven security monitoring method for serverless architectures. The method consolidates event acquisition, interaction-sequence correlation, and behavioral baseline analysis into a practical pipeline that relies on native cloud telemetry rather than invasive instrumentation of the provider platform. In this sense, the paper continues a two-year research line on serverless security: from the initial formulation of event-interaction analysis and centralized monitoring concepts to a more explicit methodological pipeline and experimental interpretation of behavioral indicators.

2. Related Work and Scientific Novelty

Recent journal studies confirm that serverless security remains an active research area and that monitoring ephemeral, event-driven workloads is still difficult. Systematic reviews of serverless security mechanisms report substantial progress at runtime, network, function, orchestration, and data layers, but also highlight persistent gaps in function-level observability and end-to-end protection of application behavior [5].

A complementary direction uses native provider monitoring to detect compromised serverless functions by modeling normal activity in cloud logs and identifying post-exploitation deviations [6]. Other journal contributions concentrate on financially motivated abuse of the pay-per-use model, known as Denial-of-Wallet attacks, through entropy-based analysis of invocation and resource features [7] or machine-learning detectors operating on serverless telemetry [8], [9]. These works demonstrate that native metrics and logs are valuable security signals, yet they often optimize for a specific threat class or for classifier performance rather than for explicit reconstruction of multi-service interaction sequences.

Surveys of serverless security and privacy further emphasize that dependency risks, limited auditability, and complex cloud-service interactions enlarge the attack surface of FaaS applications [3], [4], [10], [11]. Against this background, the scientific novelty of the present paper is an interaction-centric event-driven monitoring method that jointly uses sequence structure, invocation dynamics, and service-side metrics as a unified behavioral baseline for anomaly indication. In contrast to component-local defenses and single-threat detectors, the approach treats deviations in serverless event graphs as first-class security indicators.

3. Proposed Event-Driven Security Approach

The proposed solution formalizes serverless security monitoring as a multi-stage pipeline. The pipeline is intentionally provider-native: it consumes telemetry that is already available in managed cloud environments and therefore remains applicable when users cannot instrument the underlying FaaS runtime.

3.1. Event acquisition

The first stage aggregates events from API gateways, function invocations, message queues, storage services, and databases into a centralized log store. For each event, the pipeline retains identifiers of the triggering source, the invoked function or service, timestamps, and basic execution metrics such as duration, errors, and throttling. Centralization is required because no single component contains the full interaction path of a serverless request.

3.2. Correlation and interaction modeling

The second stage reconstructs interaction sequences. Function-to-function calls, queue-mediated hand-offs, and storage or database accesses are correlated into request-level chains. The resulting model describes not only whether a function was executed, but also how it participated in a broader event workflow. This representation makes it possible to detect structural anomalies, for example unexpected skips, loops, or newly appearing service dependencies.

3.3. Behavioral analysis and anomaly indication

The third stage builds behavioral baselines over stable operating periods and compares live event streams with those baselines. The primary indicators are invocation frequency, execution duration, and sequence structure; secondary indicators include queue backlog growth and abnormal database write latency. Deviations that persist across multiple related components are treated as candidate security incidents and can be forwarded to higher-level detection or response processes. Dependency awareness is included as a supporting risk factor: outdated or untrusted libraries in function packages may amplify the impact of anomalous execution paths.

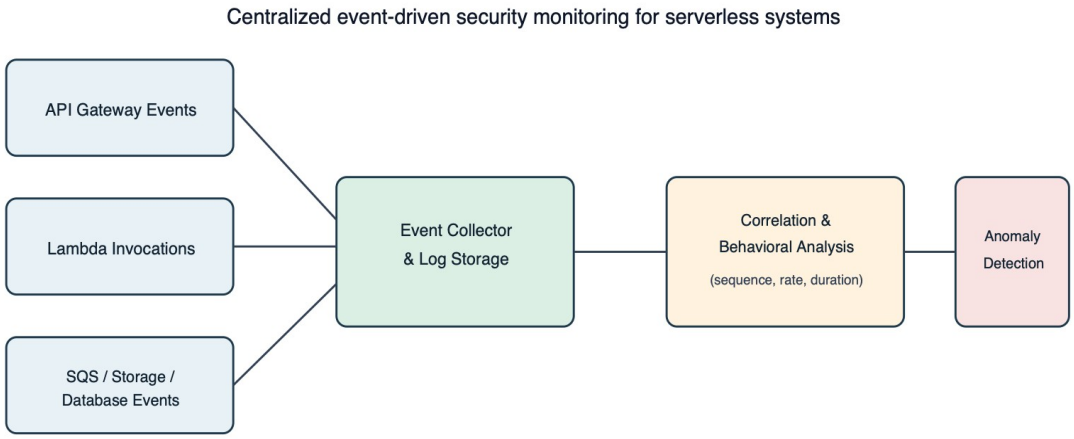


Figure 1. Centralized event-driven security monitoring pipeline for serverless architectures

Figure 1 summarizes the proposed pipeline. Events originating from API gateways, Lambda-style functions, and managed cloud services are collected into a shared store, correlated through behavioral analysis, and evaluated by an anomaly detection stage. The design separates acquisition from analysis, which allows the same monitoring logic to be reused across applications that differ in business logic but share an event-driven execution model.

4. Experimental Validation

The approach was evaluated in a controlled AWS environment. The experimental application consisted of an ingress Lambda function exposed through API Gateway, a processing function interacting with DynamoDB, an SQS queue, and a consumer function. All execution events and service metrics were observed through CloudWatch. This setting represents a typical serverless request path in which business logic is distributed across several ephemeral functions and managed services.

Under normal load, the system exhibited stable interaction structure and relatively narrow performance bands. Average function execution times remained in the range of 120–250 ms,

invocation rates varied only slightly, and the call sequence followed the designed ingress–processing–queue–consumer path. These observations define the behavioral baseline used for later comparison.

Anomalous regimes were then induced by artificially increasing the invocation frequency of selected functions by a factor of two to three and by altering the expected interaction order. In these scenarios, average execution times rose to approximately 300–450 ms, queue-related metrics indicated elevated message pressure, and atypical call sequences appeared in the correlated event stream. The experiment therefore confirms that security-relevant deviations are reflected simultaneously in temporal metrics and in the structure of event interactions.

An important practical observation is that no single metric was sufficient in isolation. Security indication became reliable when temporal anomalies co-occurred with structural changes in the correlated event chain. Because the pipeline relies on native CloudWatch telemetry, it can be deployed without privileged access to the FaaS runtime, although detection quality still depends on log completeness and baseline quality.

Table 1
Representative CloudWatch metrics in the experimental serverless chain

Component / metric	Normal regime	Anomalous regime
Typical function duration	120–250 ms	300–450 ms
Invocation frequency	Stable baseline	2–3× increase
Call-sequence structure	Ingress→Process→Queue→Consumer	Atypical / reordered paths
SQS message pressure	Low and balanced	Elevated send/receive bursts
DynamoDB write latency signal	Stable low latency	Irregular spikes

Table 1 aggregates the principal experimental indicators. Although the study does not claim a universal detection threshold for all serverless workloads, it shows that native CloudWatch telemetry is sufficient to expose abnormal operating modes when events are correlated across functions and services. This supports the use of interaction-centric behavioral monitoring as a practical security layer for FaaS applications.

5. Conclusions

This paper addressed the problem of insufficient security observability in serverless architectures whose behavior is distributed across event-triggered functions and managed cloud services. The main obtained results are as follows.

First, an event-driven security monitoring method was formulated that combines centralized event acquisition, reconstruction of function–service interaction sequences, and behavioral baseline analysis. Second, comparison with recent journal research shows that the proposed

contribution is complementary to compromised-function detectors and Denial-of-Wallet classifiers because it prioritizes cross-component interaction structure rather than a single threat signature. Third, experimental validation on AWS Lambda with CloudWatch demonstrated that anomalous regimes are observable through jointly changing invocation frequency, execution duration, and call-sequence structure. Fourth, the study confirms that meaningful security indicators can be derived from native cloud telemetry without modifying the provider-managed execution environment.

Future work will focus on automated classification of behavioral deviations under realistic production workloads and on extending the interaction model with richer dependency-risk signals, while remaining within the serverless and event-driven security setting.

References

[1] Castro, P., Ishakian, V., Muthusamy, V. and Slominski, A., "The rise of serverless computing," Communications of the ACM, vol. 62, no. 12, pp. 44–54, 2019. doi: 10.1145/3368454.

[2] Zhang, Q., Chen, M. and Chen, L., "Serverless computing: State of the art and research challenges," IEEE Access, vol. 10, pp. 115364–115379, 2022. doi: 10.1109/ACCESS.2022.3213420.

[3] Zhang, Y., Chen, X. and Li, J., "Security and privacy in serverless computing: A systematic literature review," ACM Computing Surveys, vol. 55, no. 12, Art. 240, 2023. doi: 10.1145/3579856.

[4] Li, H., Wang, Y. and Zhang, K., "Security issues and challenges in serverless computing: A survey," Journal of Cloud Computing, vol. 13, Art. 9, 2024. doi: 10.1186/s13677-024-00521-3.

[5] Escaleira, P., Cunha, V. A., Barraca, J. P., Gomes, D. and Aguiar, R. L., "A systematic review on security mechanisms for serverless computing," Cluster Computing, vol. 28, Art. 465, 2025. doi: 10.1007/s10586-025-05371-4.

[6] Ben-Shimol, L., Lavi, D., Klevansky, E., Brodt, O., Mimran, D., Elovici, Y. and Shabtai, A., "Detection of compromised functions in a serverless cloud environment," Computers & Security, vol. 150, Art. 104261, 2024. doi: 10.1016/j.cose.2024.104261.

[7] Ortega-Candel, J. M., Mora-Gimeno, F. J. and Mora, H., "Entropy-based detection of denial of wallet attacks in serverless architectures," International Journal of Information Security, vol. 25, 2026. doi: 10.1007/s10207-026-01303-4.

[8] Yalla, S. T., Batra, I. and Malik, A., "Gated recurrent unit-based detection framework for digital wallet attacks in serverless computing environments," International Journal of Networked and Distributed Computing, 2025. doi: 10.1007/s44227-025-00084-x.

[9] Renukadevi, P., Amaran, S., Vikram, A., Rao, T. P. and Ishak, M. K., "Enhancing cybersecurity through fusion of optimization with deep wavelet neural networks on denial of wallet attack detection in serverless computing," IEEE Access, vol. 13, pp. 47111–47122, 2025. doi: 10.1109/ACCESS.2025.3550735.

[10] Lin, C., Wu, Y. and Chen, J., "Security analysis of serverless architectures in cloud environments," Journal of Information Security and Applications, vol. 58, Art. 102726, 2021. doi: 10.1016/j.jisa.2021.102726.

[11] Tariq, M. and Hameed, S., "Security threats and mitigation techniques in serverless computing," Computers & Security, vol. 114, Art. 102594, 2022. doi: 10.1016/j.cose.2021.102594.

[12] Shafiei, M. and Khonsari, A., "Security challenges in serverless computing," Journal of Cloud Computing, vol. 9, Art. 55, 2020. doi: 10.1186/s13677-020-00208-0.



TRACK 3. SECURITY OF IOT DEVICES, CYBER-PHYSICAL SYSTEMS SOFTWARE, TECHNICAL SOLUTIONS FOR INFORMATION SECURITY

Multi-level protection system and access rights separation in the mobile application

Liubomyr Botseniuk[1] and Mykhailo Rizak[2]

[1] *Uzhhorod National University, Universytetska 14, 88000 Uzhhorod, Ukraine*
[2] *State University of Information and Communication Technologies, Solomianska 7, 03110 Kyiv, Ukraine*

Abstract
This paper addresses the problem of securing sensitive information in offline mobile applications. The primary focus is on mitigating local and physical threats, such as visual data interception (shoulder surfing), unauthorized screenshots, and unmonitored device access. A developed mobile financial application featuring a defense-in-depth security architecture is presented. The proposed security model comprises four layers: strict authentication (PIN and biometrics), brute-force protection with progressive lockout timers, session control with automatic background locking, and UI data leak prevention (blocking screenshots and masking previews in recent apps). A key feature of the development is the integration of a Face ID module, which is utilized not merely for binary authentication, but for user identification followed by role-based access control to system functions. The study demonstrates that the comprehensive combination of operating system mechanisms, encrypted local storage, and privacy-oriented UI logic significantly enhances the application's resilience against local data compromise.

Keywords
information security, mobile application, multi-layer security, defense-in-depth, offline storage, biometric identification, Face ID, role-based access control, UI protection.

1. Introduction

Today, a smartphone has actually become a personal repository of confidential information: from photos and correspondence to banking applications and private financial notes. That is why protecting data on the phone is no less important than protecting it online. Real risks often arise in everyday situations: peeking over your shoulder, accidental screenshot, screen recording, an open application in “recent apps” or brief access to the device without the owner. Even if the application works offline and does not transfer data to the server, the threat of leakage does not disappear - it simply shifts to local storage and interface behavior.

As part of the project, a mobile offline financial application for accounting and managing funds was created. It allows you to create multiple wallets, set the name, currency and amount, add the exchange rate and immediately see the conversion into hryvnia. The application calculates the total balance for all wallets, supports editing and deleting records, changing the order of wallets, and also has a privacy mode: amounts can be quickly masked or shown in a controlled manner - separately for each wallet or for all at once.

2. Multi-level Protection Architecture

Protection is implemented according to the principle of multi-level (defense-in-depth), when security is not reduced to a single mechanism, but is distributed over several independent layers.

Corresponding authors
liubomyr.botseniuk@uzhnu.edu.ua (L. Botseniuk); advokat.rizak@gmail.com (M. Rizak)
0000-0001-5339-9762 (L. Botseniuk); 0009-0003-9844-3271 (M. Rizak)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The first level is access authentication: login via PIN code and, if necessary, biometrics (Face ID/Touch ID) with correct processing of scenarios when the user simply cancels biometric verification.

The second level is protection against code selection: a limit of unsuccessful attempts and progressive blocking (1 second → 2 → 5 → 10 → 30...), which makes brute-force practically ineffective, while the user sees a transparent blocking timer.

The third level is session control: automatic locking when inactive and when the application returns from the background, so that financial information is not left open unattended.

The fourth level is protection against leakage through the interface: prohibiting screenshots/screen recording and protecting previews in the application switcher where supported by the platform.

A particularly important basic principle is secure storage: sensitive data and access keys are stored in the device's secure storage in an encrypted format so that even if the files are accessed, they cannot be read in plain text.

3. Face ID Module and Access Separation

An additional direction within the project was the Face ID module, which expands the classic idea of "allow/deny". In most applications, biometrics is a binary check: the user has been authenticated and has gained access. In my case, the approach is different: Face ID acts as an identification tool, that is, it allows you to not only confirm the fact of login, but also determine who exactly logged into the system. This makes it possible to bind a session to a specific person, as well as apply access delimitation depending on the role. The concept works as follows: during registration, a local profile with an identifier and a role is created for the user, and during login, biometric recognition is performed, which determines compliance with one of the registered profiles. After successful identification, the system can provide different levels of access to functions and data. Thus, biometrics becomes the basis of rights management, and not just a convenient "lock" for logging in.

A logical continuation of the application development is to deepen this mechanism: expand the system of roles and access policies, add "enhanced confirmation" scenarios for critical actions (re-identification), as well as flexibly configure privacy rules for a specific user. Also promising are modes of sharing a single device with multiple people, maintaining a secure local event log (without financial details, but with logging of login/logout facts), and strengthening the protection of local storage and backups so that even in the event of a compromise of the environment, the data remains inaccessible without keys and identity verification.

4. Conclusion

As a result, a multi-layered approach allows you to make an offline financial application not only convenient, but also resistant to typical real-world threats. It combines operating system mechanisms, thoughtful access logic, and privacy in the UI, forming a simple rule: even if one barrier fails, other layers will still reduce the risk of leakage and unauthorized access.

Declaration on Generative AI

During the preparation of this work, the author(s) used GEMINI 3.1 PRO in order to: improve grammar, and refine the styling of the author's original ideas. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

A Methodology for Building Reproducible Cybersecurity Datasets for Internet of Things Devices

Oleksandr Cherkaskyi[1]

[1] *State University of Information and Communication Technologies*

Abstract

Internet of Things intrusion datasets support comparable detection research, but many releases omit provenance, device context, labeling evidence, or leakage-resistant partitions. This paper proposes a reproducible methodology combining packet and flow records, device and gateway logs, sensor telemetry, authentication events, and configuration context. Its novelty is a temporal event graph linked to a machine-readable provenance manifest recording devices, firmware, capture conditions, scenario scripts, transformations, anonymization, and label confidence. Immutable raw evidence is separated from normalized, pseudonymized, labeled, and versioned derivatives. Partitions are disjoint by complete episode, device, time, and scenario, reducing duplicate and identity leakage. Planned validation covers metadata completeness, missingness, duplication, inter-annotator agreement, class imbalance, distribution drift, integrity, macro-F1, precision-recall area, and false alarms. A controlled testbed and held-out device families will be used; no empirical result is claimed. The contribution is an auditable protocol for reusable IoT cybersecurity benchmarks.

Keywords

Internet of Things, cybersecurity dataset, reproducibility, data provenance, intrusion detection, event graph, dataset leakage, benchmarking

1. Introduction and Problem Statement

Internet of Things systems combine sensors, gateways, mobile interfaces, cloud services, and physical processes. One incident may appear as abnormal flows, failed authentications, changed firmware, and implausible sensor behavior. Standards therefore emphasize identification, secure configuration, data protection, interface control, updating, and security-state awareness [1–3]. Data-driven detection is credible only when its evidence is inspectable and reproducible. Existing datasets enabled realistic traffic generation and device-specific botnet analysis [4–6], yet a downloadable archive is not automatically a reproducible benchmark. Reproduction also requires topology, firmware versions, clock synchronization, workloads, scenario scripts, capture filters, transformation code, label rationale, and partition rules.

Missing context can produce identical formats for different phenomena. Random row splitting may place duplicates or adjacent windows from one episode in training and test data. A model can then learn addresses, device identity, or scenario timing instead of attack behavior. This paper proposes multimodal acquisition, a temporal event graph, provenance manifests, immutable raw evidence, deterministic derivatives, and episode-, device-, time-, and scenario-disjoint partitions. It specifies a methodology and validation plan, not completed measurements.

2. Related Work and Scientific Novelty

NISTIR 8259A identifies device capabilities needed to support identification, configuration, data protection, interface restriction, secure updating, and cybersecurity-state awareness [1]. ETSI EN

Corresponding author
asherjoseph.c@gmail.com (O. Cherkaskyi)
0009-0006-3105-5217 (O. Cherkaskyi)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

303 645 supplies complementary baseline requirements for consumer IoT, while RFC 8576 describes lifecycle constraints and recurrent threats [2,3]. Together these sources clarify which security-relevant states should be observable, but they do not specify how heterogeneous observations should be captured, labeled, transformed, versioned, and partitioned as a reproducible benchmark. Existing corpora demonstrate several valid design choices.

UNSW-NB15 provides a broad network intrusion benchmark [4]; Bot-IoT combines benign and simulated malicious traffic in a dedicated IoT environment [5]; and N-BaIoT supports device-specific analysis of commercial devices affected by Mirai and BASHLITE [6]. Their scopes are intentionally different and therefore not interchangeable. Flow-only evidence may not establish whether a command changed a physical process. Device fingerprints may improve classification inside one testbed but fail under device-family holdout. Scenario-wide labels may also conceal uncertainty about attack initiation, successful compromise, observable impact, and recovery. Datasheets for Datasets promotes structured reporting of motivation, composition, collection, processing, uses, and limitations [7].

FAIR principles emphasize findability, accessibility, interoperability, and reuse [8]. Cybersecurity evidence adds a tension: unrestricted raw release may expose addresses, payload fragments, credentials, or operational weaknesses. Reproducibility must therefore coexist with minimization, pseudonymization, tiered access, and documented restrictions. The explicit scientific novelty is a four-part lifecycle. First, a temporal event graph links devices, identities, flows, commands, sensor states, attack actions, and recovery events instead of leaving sources as unrelated tables. Second, every released object receives a machine-readable manifest recording firmware, configuration, clock state, scenario version, transformation history, label source, confidence, and permitted use.

Third, immutable evidence is separated from derivatives generated by deterministic scripts and bound to inputs through cryptographic digests and semantic versions. Fourth, complete episodes are partitioned jointly by device, time, and scenario, enabling distinct tests of interpolation, temporal transfer, and cross-device generalization. Quality remains multidimensional: completeness, agreement, duplicates, missingness, imbalance, drift, integrity, and baseline behavior are disclosed separately. An additional distinction is the treatment of label confidence as reportable evidence rather than a value erased during adjudication. This supports later reinterpretation when threat taxonomies change. The contribution is this auditable methodology and its prospective evaluation criteria, not an unperformed claim of superior detection performance.

2.1. Formal Model

$$Cmeta = observed_required_fields / total_required_fields \quad (1)$$

where Cmeta is completeness; observed_required_fields counts populated mandatory fields; total_required_fields equals object count multiplied by mandatory fields per object. Quantifies provenance metadata coverage.

$$kappa = (po - pe) / (1 - pe) \quad (2)$$

where po is observed independent-annotator agreement; pe is chance agreement estimated from annotators' marginal label distributions. Quantifies pre-adjudication labeling reliability.

$$MacroF1 = (1/K) * \sum_k(2 * Pk * Rk / (Pk + Rk)) \quad (3)$$

where K is class count; Pk and Rk are precision and recall for class k; classes with undefined denominators contribute zero. Summarizes class-balanced baseline detection performance.

3. Proposed Solution and Method

The method begins with a threat model, inventory, and capture plan. For each controlled device, the inventory records family, hardware revision, firmware, services, network position, update state, and expected physical behavior. Benign workloads are separated from reconnaissance, credential abuse, spoofing, malicious-update attempts, denial of service, and process manipulation. A controller records synchronized markers for every action, effect, and recovery. Packet captures and flow summaries are paired with gateway, device, application, authentication, configuration, and sensor records. Sources have stable identifiers and declared sampling policies. Raw objects are write-once and hashed.

A deterministic pipeline verifies integrity, normalizes schemas and time, handles duplicates under documented rules, minimizes payloads, pseudonymizes identifiers, and records software versions, parameters, transformations, and rejected objects. Two reviewers independently determine whether each intended action started, reached its target, succeeded, caused impact, or remained uncertain. Disagreements persist until adjudication, while pre-adjudication agreement is reported with κ. Bounded time windows, stable pseudonyms, and typed relations join observations into the temporal graph. The manifest and datasheet state uses, exclusions, privacy controls, and blind spots. Each semantic release contains schemas, manifests, hashes, code, labels, and fixed partitions.

Complete episodes enter training, validation, or testing; stricter profiles hold out later periods and device families. Learned preprocessing uses training data only. Equations define metadata completeness, agreement, and macro-F1. Corrections create a new version and change log.

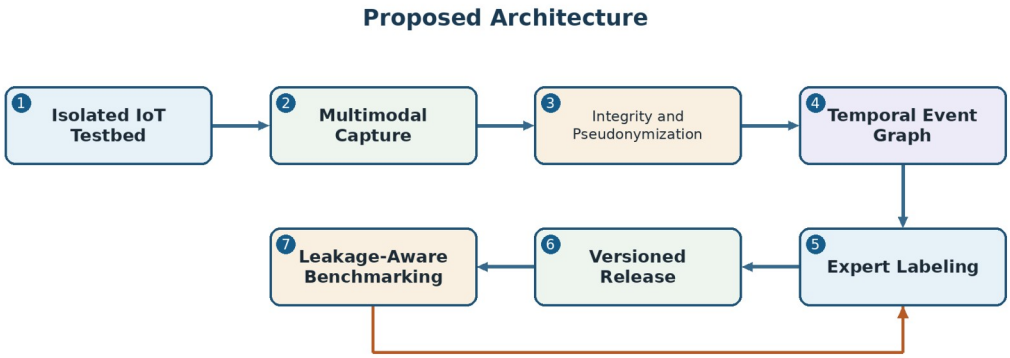


Figure 1: Reproducible lifecycle for an IoT cybersecurity dataset

4. Evaluation Design

Validation is a staged audit. Automated checks cover schemas, hashes, manifest links, timestamp order, pseudonym consistency, deterministic transformations, missingness, and duplicates. A second reviewer labels samples from every scenario and device family; κ is calculated before adjudication. Split integrity is tested using episode identifiers, hashes, time overlap, and device pseudonyms. Frozen linear, tree-based, and one-class baselines use validation-only tuning. Planned measures include per-class precision and recall, macro-F1, precision-recall area, false alarms per benign hour, and episode latency. Identifier-only probes expose leakage. Whole episodes support uncertainty intervals. Table 1 contains illustrative planning values, not observations.

Ablations compare network-only and graph inputs, random and episode-disjoint splits, and known and held-out devices. Reproduction requires identical derivative hashes and partitions.

Table 1
Proposed validation matrix (illustrative study design; no empirical results are reported)

Scenario	Proposed capture	Proposed repetitions	Holdout profile	Primary planned checks
Benign daily workload	24 h per device family	3	Later time window	Missingness, drift, false alarms/hour
Reconnaissance	60 min per run	5	Unseen device model	Episode recall, detection latency
Credential abuse	90 min per run	5	Unseen account set	Precision, recall, label agreement
Message spoofing	60 min per run	5	Unseen payload pattern	Macro-F1, PR-AUC, graph ablation
Denial of service	30 min per run	5	Unseen rate profile	Recall, time to detection, missingness
Physical-process manipulation	90 min per run	4	Unseen device family	Impact-label agreement, episode recall

Note. Values are illustrative design parameters, not measured results.

5. Discussion

The proposed lifecycle treats an IoT cybersecurity dataset as a versioned research instrument rather than a static table. Its principal advantage is traceability: users can identify the device, firmware state, scenario action, transformation, label decision, and partition rule supporting an analysis. The event graph also connects a malicious command with device-state and physical effects that network-only rows may omit. These benefits require immutable evidence and retained transformation histories. Operational tradeoffs remain. Multimodal acquisition requires synchronized clocks, controlled workloads, storage, protected pseudonym mappings, and expert review.

Some devices expose few logs; encryption limits inspection; clock drift complicates joins; and distinctive traffic can remain identifiable after pseudonymization. Tiered access protects sensitive captures but may limit direct replication, making independent scenario execution an important alternative. Validity depends on the testbed. Scripted attacks improve safety and repeatability but may not reproduce adversarial adaptation, background diversity, or firmware change. Episode-disjoint partitions reduce leakage without proving deployment generalization, and baseline models expose shortcuts without certifying operational readiness.

6. Conclusions

Datasheets should state excluded devices, inaccessible evidence, uncertainty, intended uses, licensing, disclosure procedures, retention, and versioned corrections. Governance must define who may authorize raw-data access and how misuse or correction requests are handled. In conclusion, the paper defines a reproducible lifecycle joining multimodal evidence, explicit provenance, label uncertainty, privacy-aware release layers, and leakage-resistant partitions. Its scientific output is the method, formal quality measures, architecture, and validation protocol; no empirical advantage is asserted.

Future work should implement the testbed, preregister scripts, evaluate reviewer agreement, publish deterministic environments, and test later-time and unseen-device holdouts. A second laboratory should reproduce one scenario with a different device family and report successes and

failures. Acceptance should depend on transparent audit evidence, reproducibility, and disclosed limitations, not on a predetermined detection score.

Declaration on Generative AI

Generative AI tools were employed exclusively for language editing, grammatical refinement, and visual figure generation. All scientific content, analysis, and conclusions were authored, verified, and finalized by the author(s).

References

[1] Fagan, M., Megas, K. N., Scarfone, K., and Smith, M. IoT Device Cybersecurity Capability Core Baseline. NISTIR 8259A, National Institute of Standards and Technology, Gaithersburg, MD, 2020. https://doi.org/10.6028/NIST.IR.8259A.

[2] European Telecommunications Standards Institute. Cyber Security for Consumer Internet of Things: Baseline Requirements. ETSI EN 303 645 V2.1.1, Sophia Antipolis, France, June 2020.

[3] Garcia-Morchon, O., Kumar, S., and Sethi, M. Internet of Things (IoT) Security: State of the Art and Challenges. RFC 8576, Internet Engineering Task Force, April 2019. https://doi.org/10.17487/RFC8576.

[4] Moustafa, N., and Slay, J. UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set). In: 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, Australia, pp. 1–6, 2015. https://doi.org/10.1109/MilCIS.2015.7348942.

[5] Koroniotis, N., Moustafa, N., Sitnikova, E., and Turnbull, B. Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset. Future Generation Computer Systems, 100, 779–796, 2019. https://doi.org/10.1016/j.future.2019.05.041.

[6] Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Breitenbacher, D., Shabtai, A., and Elovici, Y. N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders. IEEE Pervasive Computing, 17(3), 12–22, 2018. https://doi.org/10.1109/MPRV.2018.03367731.

[7] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for Datasets. Communications of the ACM, 64(12), 86–92, 2021. https://doi.org/10.1145/3458723.

[8] Wilkinson, M. D., et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data, 3, 160018, 2016. https://doi.org/10.1038/sdata.2016.18.

Block Data Encryption Using Polynomial Modified Fibonacci Matrices

Pavlo Grytsiuk[1] and Yurii Hrytsiuk[1]

[1] *Lviv Polytechnic National University, Stepan Bandera Street 12, 79013 Lviv, Ukraine*

Abstract

Various approaches to generating sequences of Fibonacci polynomial matrices have been substantiated, and the expediency of their application for block data encryption has been proven. Shortcomings in the traditional approach to forming the structure of elements of such matrices—primarily regarding the number (k) of its distinct elements, which are Fibonacci polynomials of degree no higher than (n—1) — have been identified. For Fibonacci matrices of any order (m), there are only k = 3 such distinct Fibonacci polynomials, the structure of which depends solely on the sequence index (n) of the Fibonacci polynomial matrix. Such a small number of elements is not only uninformative and transparent to a cryptanalyst, but also vulnerable to cryptanalysis. The structure of elements in Fibonacci polynomial matrices has been modified, with their quantity now depending on the order of the matrix (m) and being equal to $k_m = 2^{m-1}$. A method for generating a sequence of modified Fibonacci polynomial matrices using a recurrent matrix relation has been developed. Algorithms for constructing sequences of modified Fibonacci polynomial matrices of orders 2 through 5 are presented, whose elements are Fibonacci polynomials of index no higher than (2(m—2)+n). This enabled the analysis of not only the structural characteristics of their construction, but also a thorough understanding of the corresponding procedures for calculating their determinants and inverse matrices. Software has been developed that enables the generation of sequences of modified Fibonacci polynomial matrices of the mth order, as well as the calculation of their determinants and the computation of inverse polynomial matrices of the same order.

Keywords

The Fibonacci polynomial sequence, a recursive matrix relation, the inverse Fibonacci polynomial matrix, an encrypted message, a matrix-based data encryption method

1. Introduction

Advances in information technology have contributed to the application of Fibonacci polynomial matrices and their various modifications across various fields of knowledge [9], including block data encryption primarily using the matrix method [6]. Although a considerable number of studies on the generation of Fibonacci polynomial matrix sequences — whose elements consist of Fibonacci polynomials and their diverse analogues — have occasionally appeared in open publication over the past few decades, they remain significantly less explored than so-called Fibonacci matrices composed of various Fibonacci numbers. The authors of these and many other studies suggest that using Fibonacci polynomial matrix sequences as a standalone data encryption procedure for transmission over communication channels is impractical due to the small number of distinct elements in such matrices, as well as the low cryptographic strength of the encryption algorithm.

However, even a superficial analysis of Fibonacci polynomial matrix sequences of orders 2 through 5, presented in study [5], reveals that employing the traditional element structure of such matrices is inadvisable due to the small quantity (k) of distinct Fibonacci polynomials of degree no higher than (n—1). For instance, for Fibonacci matrices of any order (m), there are only k = 3 such distinct Fibonacci polynomials, whose degrees depend solely on the sequence index (n) of the

Corresponding authors

pavlo.y.hrytsiuk@lpnu.ua (P. Grytsiuk); yurii.i.hrytsiuk@lpnu.ua (Yu. Hrytsiuk)

0009-0003-5409-2043 (P. Grytsiuk); 0000-0001-8183-3466 (Yu. Hrytsiuk)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Fibonacci polynomial matrix — specifically ranging from degree (n—3) to (n—1). Such a limited number of distinct elements and the minor variation in polynomial degrees render the structure not only uninformative and transparent to a cryptanalyst, but also vulnerable to cryptanalysis. Consequently, there is a need for additional research into the specific characteristics of generating Fibonacci polynomial matrix sequences and modifying the structure of their elements. This modification should ensure the ability to compute both their determinants and inverse matrices, while enabling the identification of specific features for their application strictly as a local block data encryption method.

2. Materials and methods

2.1. A method for generating a sequence of Fibonacci polynomials

The following recursive relation is used to generate the Fibonacci sequence [6]:

$$F^{(n+1)}(x) = xF^{(n)}(x) + F^{(n-1)}(x)\,, \tag{1}$$

where $F^{(n+1)}(x)$, $F^{(n)}(x)$ and $F^{(n-1)}(x)$ are, respectively, the next, current and previous terms of the Fibonacci polynomial. Applying the recursive relation (1), we obtain a sequence of Fibonacci polynomials, which takes the following form:

$$\begin{aligned}
&F^{(0)}(x) = 0; \quad F^{(1)}(x) = 1; \quad F^{(2)}(x) = x; \quad F^{(3)}(x) = x^2 + 1;\\
&F^{(4)}(x) = x^3 + 2x; \quad F^{(5)}(x) = x^4 + 3x^2 + 1; \quad F^{(6)}(x) = x^5 + 4x^3 + 3x;\\
&F^{(7)}(x) = x^6 + 5x^4 + 6x^2 + 1; \quad F^{(8)}(x) = x^7 + 6x^5 + 10x^3 + 4x;\\
&F^{(9)}(x) = x^8 + 7x^6 + 15x^4 + 10x^2 + 1; \quad F^{(10)}(x) = x^9 + 8x^7 + 21x^5 + 20x^3 + 5x;\\
&\cdots \quad \cdots \quad \cdots \quad \cdots\\
&F^{(15)}(x) = x^{14} + 13x^{12} + 66x^{10} + 165x^8 + 210x^6 + 126x^4 + 28x^2 + 1.
\end{aligned} \tag{2}$$

The method for generating the Fibonacci polynomial sequence ‘forwards’ involves using the recursive relation (1), according to which the next polynomial is formed by multiplying the variable x successively by the terms of the current polynomial, after which like terms are grouped with those of the previous polynomial. Fibonacci polynomial sequences are a natural extension of the corresponding Fibonacci number sequences; therefore, many of their properties can be proved directly [3, 7, 10].

2.2. The mechanism for generating a sequence of modified Fibonacci polynomial matrices

A 2×2 Fibonacci polynomial matrix takes the following form [9]:

$$\bar{\bar{Q}}_2^{(n+1)}(x) = \begin{bmatrix} F^{(n+1)}(x) & F^{(n)}(x) \\ F^{(n)}(x) & F^{(n-1)}(x) \end{bmatrix} \text{for} \forall n \in \square \tag{3}$$

and its determinant is obtained from Cassini’s identity [6]

$$\det\left(\bar{\bar{Q}}_2^{(n+1)}(x)\right) = F^{(n+1)}(x)F^{(n-1)}(x) - (F^{(n)}(x))^2 = (-1)^{n+1} \tag{4}$$

Fibonacci polynomial matrices of orders 2 to 5, as well as of order m, take the following form [8]:

$$\bar{\bar{Q}}_2^{(2)}(x) = \begin{bmatrix} x & 1 \\ 1 & 0 \end{bmatrix}; \bar{\bar{Q}}_3^{(2)}(x) = \begin{bmatrix} x & 1 & 0 \\ 0 & x & 1 \\ 0 & 1 & 0 \end{bmatrix}; \bar{\bar{Q}}_4^{(2)}(x) = \begin{bmatrix} x & 1 & 0 & 0 \\ 0 & x & 1 & 0 \\ 0 & 0 & x & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}; \quad \bar{\bar{Q}}_5^{(2)}(x) = \begin{bmatrix} x & 1 & 0 & 0 & 0 \\ 0 & x & 1 & 0 & 0 \\ 0 & 0 & x & 1 & 0 \\ 0 & 0 & 0 & x & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}; \bar{\bar{Q}}_m^{(2)}(x) = \begin{bmatrix} x & 1 & 0 & 0 & \cdots & 0 \\ 0 & x & 1 & 0 & \cdots & 0 \\ 0 & 0 & x & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & x & 1 \\ 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix}. \tag{5}$$

To generate the nth sequence of Fibonacci polynomial matrices, a recursive matrix relation is used, analogous to relation (1), which in our case takes the following form:

$$\overline{\overline{Q}}_m^{(n+1)}(x) = x\overline{\overline{Q}}_m^{(n)}(x) + \overline{\overline{Q}}_m^{(n-1)}(x), \forall n \geq 2 \tag{6}$$

where $\overline{\overline{Q}}_m^{(n+1)}(x), \overline{\overline{Q}}_m^{(n)}(x)$ and $\overline{\overline{Q}}_m^{(n-1)}(x)$ denotes, respectively, the next, current and previous forms of the polynomial $\overline{\overline{Q}}_m^{(\square)}(x)$ -matrices of Fibonacci.

As can be seen from matrix expressions (5), as well as from the results of the analysis of Fibonacci polynomial matrix sequences of orders 2 through 5 presented in study [5], the traditional approach to forming the structure of elements of such matrices exhibits certain shortcomings. One of these shortcomings concerns the number (k) of its distinct elements, which are Fibonacci polynomials of degree no higher than (n—1). For instance, for Fibonacci matrices of any order (m), there are only k = 3 such distinct Fibonacci polynomials, whose degree depends solely on the nth sequence of the Fibonacci polynomial matrix (see matrix expressions (7), (19), (20), or (21) in study [5]). Such a small number of elements is not only uninformative and transparent to a cryptanalyst, but also vulnerable to cryptanalysis.

In view of these circumstances, we consider it advisable to slightly modify the structure of the elements of Fibonacci polynomial matrices, the number of which should primarily depend on their order (m) and the sequence index (n) of the polynomial matrix. Furthermore, for a better understanding of the material presented below, we will use the indices of Fibonacci polynomials from mathematical expressions (3) rather than their degrees — for example, polynomials $F^{(3)}(x)$ or $F^{(5)}(x)$, whose indices (n) are 3 and 5, respectively. Thus, for a Fibonacci matrix of order m = 2, its element structure will remain unchanged (7); that is, the number of its distinct elements, which are Fibonacci polynomials of index no higher than n, will be $k_2 = m + 1 = 2 + 1 = 3$.

$$\begin{gathered}
\overline{\overline{Q}}_2^{(2)}(x) = \begin{bmatrix} x & 1 \\ 1 & 0 \end{bmatrix}; \overline{\overline{Q}}_3^{(2)}(x) = \begin{bmatrix} x^3+2x & x^2+1 & x \\ 0 & x & 1 \\ 0 & 1 & 0 \end{bmatrix}_3; \overline{\overline{Q}}_4^{(2)}(x) = \begin{bmatrix} x^5+4x^3+3x & x^4+3x^2+1 & x^3+2x & x^2+1 \\ 0 & x^3+2x & x^2+1 & x \\ 0 & 0 & x & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}_4; \\
\overline{\overline{Q}}_5^{(2)}(x) = \begin{bmatrix} x^7+6x^5+10x^3+4x & x^6+5x^4+6x^2+1 & x^5+4x^3+3x & x^4+3x^2+1 & x^3+2x \\ 0 & x^5+4x^3+3x & x^4+3x^2+1 & x^3+2x & x^2+1 \\ 0 & 0 & x^3+2x & x^2+1 & x \\ 0 & 0 & 0 & x & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}_5; \\
\overline{\overline{Q}}_m^{(2)}(x) = \begin{bmatrix} (m+3) & (m+2) & (m+1) & (m) & \cdots & (m-1) \\ 0 & (m+1) & (m) & (m-1) & \cdots & (m-2) \\ 0 & 0 & (m-1) & (m-2) & \cdots & (m-3) \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & x & 1 \\ 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix}_m.
\end{gathered} \tag{7}$$

$$\overline{\overline{Q}}_2^{(2)}(x) = \begin{bmatrix} F^{(2)}(x) & F^{(1)}(x) \\ F^{(1)}(x) & F^{(0)}(x) \end{bmatrix}; \overline{\overline{Q}}_3^{(2)}(x) = \begin{bmatrix} F^{(4)}(x) & F^{(3)}(x) & F^{(2)}(x) \\ 0 & F^{(2)}(x) & F^{(1)}(x) \\ 0 & F^{(1)}(x) & F^{(0)}(x) \end{bmatrix}_{3\times3}; \overline{\overline{Q}}_4^{(2)}(x) = \begin{bmatrix} F^{(6)}(x) & F^{(5)}(x) & F^{(4)}(x) & F^{(3)}(x) \\ 0 & F^{(4)}(x) & F^{(3)}(x) & F^{(2)}(x) \\ 0 & 0 & F^{(2)}(x) & F^{(1)}(x) \\ 0 & 0 & F^{(1)}(x) & F^{(0)}(x) \end{bmatrix}_{4\times4}; \tag{8}$$

$$\overline{\overline{Q}}_5^{(2)}(x) = \begin{bmatrix} F^{(8)}(x) & F^{(7)}(x) & F^{(6)}(x) & F^{(5)}(x) & F^{(4)}(x) \\ 0 & F^{(6)}(x) & F^{(5)}(x) & F^{(4)}(x) & F^{(3)}(x) \\ 0 & 0 & F^{(4)}(x) & F^{(3)}(x) & F^{(2)}(x) \\ 0 & 0 & 0 & F^{(2)}(x) & F^{(1)}(x) \\ 0 & 0 & 0 & F^{(1)}(x) & F^{(0)}(x) \end{bmatrix}_5; \overline{\overline{Q}}_m^{(2)}(x) = \begin{bmatrix} F^{(m+3)}(x) & F^{(m+2)}(x) & F^{(m+1)}(x) & F^{(m)}(x) & \cdots & F^{(m-1)}(x) \\ 0 & F^{(m+1)}(x) & F^{(m)}(x) & F^{(m-1)}(x) & \cdots & F^{(m-2)}(x) \\ 0 & 0 & F^{(m-1)}(x) & F^{(m-2)}(x) & \cdots & F^{(m-3)}(x) \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & F^{(2)}(x) & F^{(1)}(x) \\ 0 & 0 & \cdots & 0 & F^{(1)}(x) & F^{(0)}(x) \end{bmatrix}_m.$$

At the same time, for a Fibonacci matrix of order m = 3, we will attempt to modify the element structure slightly to increase the number of polynomial types up to index 4. This means that the number of distinct elements in the matrix, which consist of Fibonacci polynomials, will now be k_3 = 3 + 2 = 5. Similarly, we modify the element structure of Fibonacci matrices of orders 4 and 5 to

increase the number of their Fibonacci polynomial types up to indices 4 and 5, respectively; that is, the number of distinct elements in these matrices will be $k_4 = 4 + 3 = 7$ and $k_5 = 5 + 4 = 9$, respectively. In expressions (7) for the Fibonacci matrix, the indices of their Fibonacci polynomials are indicated in parentheses above the elements, with their total quantity amounting to $k_m = 2m—1$.

The indices of the Fibonacci polynomials specified in matrix expressions (7) pertain solely to the $n = 2$ sequence of modified mth order Fibonacci polynomial matrices. For higher sequence indices ($n \geq 3$) of the modified Fibonacci matrices, taking into account their order (m), the indices of their Fibonacci polynomials will assume somewhat different values. To determine these Fibonacci polynomial indices, let us consider matrix expressions (16), which display the elements of modified mth order Fibonacci matrices in the form of identifiers of the corresponding Fibonacci polynomials for the $n = 2$ sequence of polynomial matrices.

2.3. The mechanism for generating a sequence of modified 3rd-order Fibonacci polynomial matrices

It can be seen from the matrix expressions (9) given above that the elements of the nth modified Fibonacci polynomial matrix are the corresponding Fibonacci polynomials numbered from (n–2) to (2(m–2)+n) in their sequence (3).

$$\bar{\bar{Q}}_3^{(n+1)}(x) = x\bar{\bar{Q}}_3^{(n)}(x) + \bar{\bar{Q}}_3^{(n-1)}(x);\ \det(\bar{\bar{Q}}_3^{(n)}(x)) = (-1)^{n+1}F^{(n+2)}(x);\ \bar{\bar{Q}}_3^{(n)}(x) \times \bar{\bar{Q}}_3^{-(n)}(x) = \bar{\bar{I}}_3;$$

$$\bar{\bar{Q}}_3^{(0)}(x) = \bar{\bar{Q}}_3^{(2)}(x) - x\bar{\bar{Q}}_3^{(1)}(x) = \begin{bmatrix} x & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & -x \end{bmatrix};\ \det(\bar{\bar{Q}}_3^{(0)}(x)) = -x;\ \bar{\bar{Q}}_3^{-(0)}(x) = \begin{bmatrix} 1/x & -1 & -1/x \\ 0 & x & 1 \\ 0 & 1 & 0 \end{bmatrix};$$

$$\bar{\bar{Q}}_3^{(1)}(x) = \bar{\bar{Q}}_3^{(3)}(x) - x\bar{\bar{Q}}_3^{(2)}(x) = \begin{bmatrix} x^2+1 & x & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix};\ \det(\bar{\bar{Q}}_3^{(1)}(x)) = x^2 + 1;\ \bar{\bar{Q}}_3^{-(1)}(x) = \begin{bmatrix} \frac{1}{x^2+1} & \frac{-x}{x^2+1} & \frac{-1}{x^2+1} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix};$$

$$\bar{\bar{Q}}_3^{(2)}(x) = x\bar{\bar{Q}}_3^{(1)}(x) + \bar{\bar{Q}}_3^{(0)}(x) = \begin{bmatrix} x^3+2x & x^2+1 & x \\ 0 & x & 1 \\ 0 & 1 & 0 \end{bmatrix};\ \det(\bar{\bar{Q}}_3^{(2)}(x)) = -(x^3 + 2x);\ \bar{\bar{Q}}_3^{-(2)}(x) = \begin{bmatrix} \frac{1}{x^3+2x} & \frac{-x}{x^3+2x} & \frac{-1}{x^3+2x} \\ 0 & 0 & 1 \\ 0 & 1 & -x \end{bmatrix};$$

$$\bar{\bar{Q}}_3^{(3)}(x) = x\bar{\bar{Q}}_3^{(2)}(x) + \bar{\bar{Q}}_3^{(1)}(x) = \begin{bmatrix} x^4+3x^2+1 & x^3+2x & x^2+1 \\ 0 & x^2+1 & x \\ 0 & x & 1 \end{bmatrix};\ \bar{\bar{Q}}_3^{-(3)}(x) = \begin{bmatrix} \frac{1}{\det(\bar{\bar{Q}}_3^{(3)}(x))} & \frac{-x}{\det(\bar{\bar{Q}}_3^{(3)}(x))} & \frac{-1}{\det(\bar{\bar{Q}}_3^{(3)}(x))} \\ 0 & 1 & -x \\ 0 & -x & x^2+1 \end{bmatrix}; \quad (9)$$

$$\det(\bar{\bar{Q}}_3^{(3)}(x)) = x^4 + 3x^2 + 1;$$

$$\bar{\bar{Q}}_3^{(4)}(x) = x\bar{\bar{Q}}_3^{(3)}(x) + \bar{\bar{Q}}_3^{(2)}(x) = \begin{bmatrix} x^5+4x^3+3x & x^4+3x^2+1 & x^3+2x \\ 0 & x^3+2x & x^2+1 \\ 0 & x^2+1 & x \end{bmatrix};\ \bar{\bar{Q}}_3^{-(4)}(x) = \begin{bmatrix} \frac{-1}{\det(\bar{\bar{Q}}_3^{(4)}(x))} & \frac{x}{\det(\bar{\bar{Q}}_3^{(4)}(x))} & \frac{1}{\det(\bar{\bar{Q}}_3^{(4)}(x))} \\ 0 & -x & x^2+1 \\ 0 & x^2+1 & -(x^3+2x) \end{bmatrix};$$

$$\det(\bar{\bar{Q}}_3^{(4)}(x)) = -(x^5 + 4x^3 + 3x);$$

$$\bar{\bar{Q}}_3^{(5)}(x) = x\bar{\bar{Q}}_3^{(4)}(x) + \bar{\bar{Q}}_3^{(3)}(x) = \begin{bmatrix} x^6+5x^4+6x^2+1 & x^5+4x^3+3x & x^4+3x^2+1 \\ 0 & x^4+3x^2+1 & x^3+2x \\ 0 & x^3+2x & x^2+1 \end{bmatrix};\ \bar{\bar{Q}}_3^{-(5)}(x) = \begin{bmatrix} \frac{1}{\det(\bar{\bar{Q}}_3^{(5)}(x))} & \frac{-x}{\det(\bar{\bar{Q}}_3^{(5)}(x))} & \frac{-1}{\det(\bar{\bar{Q}}_3^{(5)}(x))} \\ 0 & x^2+1 & -(x^3+2x) \\ 0 & -(x^3+2x) & x^4+3x^2+1 \end{bmatrix};$$

$$\det(\bar{\bar{Q}}_3^{(5)}(x)) = x^6 + 5x^4 + 6x^2 + 1.$$

Since we are dealing with Fibonacci matrices of order 3 and higher, special mathematical methods must be employed to compute their determinants [4]. Examples include recursive determinant computation through order reduction, known as Laplace expansion, as well as the Bareiss algorithm or the Leibniz formula. As a reminder, Erwin Bareiss's algorithm is an algorithm for computing the determinant of a matrix with integer elements by reducing it to row echelon form using strictly integer arithmetic. Here, integer elements encompass both numerical and symbolic matrix values, including mathematical expressions. Any division performed by the

algorithm is guaranteed to be an exact division without a remainder. The algorithm can also be used to compute the determinant of a matrix with (approximate) real elements, which eliminates rounding errors, except for those already present in the input data.

To compute the inverse of a Fibonacci polynomial matrix of order 3 and higher, appropriate mathematical methods must likewise be used [4], such as the Montante method, Gauss-Jordan elimination, or the adjugate matrix approach. For instance, the Montante method is a linear algebra method designed for solving systems of linear equations, as well as finding inverse matrices and determinants. The method is named after its discoverer, René Mario Montante Pardo. Its main feature is that it operates using exclusively integer arithmetic for integer or symbolic matrices, which enables accurate results to be obtained in computer implementations.

3. A method for encrypting block data using Fibonacci polynomial matrices

Let us consider the implementation details of the matrix data encryption method using *m*th order Fibonacci polynomial matrices [2, 7]. To utilize this method, the initial message must be presented in the form of a square *m*th order matrix, referred to as the message matrix. There are no restrictions on how such a message is represented; thus, the choice of arranging the message elements in the matrix — either row by row or column by column — is left to the user. For instance, the input message "*Back_to_the_drawing_board*" can be represented as a 4th-order matrix by filling it element by element by rows, while simultaneously indicating their corresponding numerical values according to the ASCII character code table:

$$\bar{\bar{M}} = \begin{bmatrix} B & a & c & k & _ \\ t & o & _ & t & h \\ e & _ & d & r & a \\ w & i & n & g & _ \\ b & o & a & r & d \end{bmatrix} \Rightarrow \begin{bmatrix} 66 & 97 & 99 & 107 & 95 \\ 116 & 111 & 95 & 116 & 104 \\ 101 & 95 & 100 & 114 & 97 \\ 119 & 105 & 110 & 103 & 95 \\ 98 & 111 & 97 & 114 & 100 \end{bmatrix}.$$

Once the sender and receiver have agreed on an integer x ≠ 0 and an n-th Fibonacci polynomial matrix of order m, the corresponding data encryption matrix is selected from equation (9) given above. This matrix must then be multiplied on the right-hand side by the message matrix $\bar{\bar{M}}$ to obtain the encrypted message matrix $\bar{\bar{E}}$. For example, for m = 5 and n = 3, from the matrix expression (24) we have:

$$\bar{\bar{Q}}_5^{(3)}(x) = x\bar{\bar{Q}}_5^{(2)}(x) + \bar{\bar{Q}}_5^{(1)}(x) = \begin{bmatrix} F^{(9)}(x) & F^{(8)}(x) & F^{(7)}(x) & F^{(6)}(x) & F^{(5)}(x) \\ 0 & F^{(7)}(x) & F^{(6)}(x) & F^{(5)}(x) & F^{(4)}(x) \\ 0 & 0 & F^{(5)}(x) & F^{(4)}(x) & F^{(3)}(x) \\ 0 & 0 & 0 & F^{(3)}(x) & F^{(2)}(x) \\ 0 & 0 & 0 & F^{(2)}(x) & F^{(1)}(x) \end{bmatrix};$$

$$\bar{\bar{Q}}_5^{-(3)}(x) = \begin{bmatrix} \frac{q'_{11}}{\det(\bar{\bar{Q}}_5^{(3)}(x))} & \frac{-q'_{12}}{\det(\bar{\bar{Q}}_5^{(3)}(x))} & \frac{-1}{\det(\bar{\bar{Q}}_5^{(3)}(x))} & \frac{q'_{14}}{\det(\bar{\bar{Q}}_5^{(3)}(x))} & \frac{-q'_{15}}{\det(\bar{\bar{Q}}_5^{(3)}(x))} \\ 0 & \frac{F^{(5)}(x)}{q''_{22}} & \frac{-F^{(6)}(x)}{q''_{23}} & \frac{-F^{(3)}(x)}{q''_{24}} & \frac{F^{(4)}(x)}{q''_{25}} \\ 0 & 0 & \frac{F^{(1)}(x)}{F^{(5)}(x)} & \frac{-F^{(2)}(x)}{F^{(5)}(x)} & \frac{-F^{(1)}(x)}{F^{(5)}(x)} \\ 0 & 0 & 0 & F^{(1)}(x) & -F^{(2)}(x) \\ 0 & 0 & 0 & -F^{(2)}(x) & F^{(3)}(x) \end{bmatrix};$$

$$\det(\bar{\bar{Q}}_5^{(3)}(x)) = x^{18} + 15x^{16} + 93x^{14} + 308x^{12} + 588x^{10} + 652x^8 + 404x^6 + 129x^4 + 19x^2 + 1;$$

$$q'_{12} = x^{11} + 9x^9 + 29x^7 + 40x^5 + 22x^3 + 4x;\ q'_{14} = x^7 + 5x^5 + 7x^3 + 3x;\ q'_{15} = x^8 + 6x^6 + 11x^4 + 6x^2;$$

$$q'_{11} = q''_{22} = q''_{23} = q''_{24} = q''_{25} = x^{10} + 8x^8 + 22x^6 + 24x^4 + 9x^2 + 1;$$

Now, by substituting the value x = 2 into this matrix and carrying out the relevant calculations, we obtain:

$$\overline{\overline{Q}}_5^{(3)}(2)=\begin{bmatrix}509626 & 408 & 169 & 70 & 29\\ 0 & 169 & 70 & 29 & 12\\ 0 & 0 & 29 & 12 & 5\\ 0 & 0 & 0 & 5 & 2\\ 0 & 0 & 0 & 2 & 1\end{bmatrix};\ \overline{\overline{E}}=\overline{\overline{Q}}_5^{(3)}(2)\times\overline{\overline{M}}=\begin{bmatrix}33710885 & 49505634 & 50519147 & 54607092 & 48482845\\ 31301 & 29786 & 27409 & 31939 & 28321\\ 4847 & 4570 & 4705 & 5112 & 4453\\ 791 & 747 & 744 & 743 & 675\\ 336 & 321 & 317 & 320 & 290\end{bmatrix}.$$

Assuming that the transmitted sequence of numbers has been received without errors, then, by multiplying the inverse Fibonacci matrix, taken from the matrix expression (24), by the matrix of the encrypted message $\overline{\overline{E}}$, given above, the original message matrix is obtained, namely:

$$\overline{\overline{Q}}_4^{-(3)}(2)=\begin{bmatrix}0{,}0000019622 & -0{,}00000474 & -0{,}00000000040 & 0{,}0000001401 & -0{,}000000336\\ 0 & 0{,}00591716 & -0{,}0142827994 & -0{,}001020200 & 0{,}002448480\\ 0 & 0 & 0{,}0344827586 & -0{,}068965517 & -0{,}034482759\\ 0 & 0 & 0 & 1 & -2\\ 0 & 0 & 0 & -2 & 5\end{bmatrix};$$

$$\overline{\overline{M}}=\overline{\overline{Q}}_5^{-(3)}(2)\times\overline{\overline{E}}=\begin{bmatrix}66 & 97 & 99 & 107 & 95\\ 116 & 111 & 95 & 116 & 104\\ 101 & 95 & 100 & 114 & 97\\ 119 & 105 & 110 & 103 & 95\\ 98 & 111 & 97 & 114 & 100\end{bmatrix}\Rightarrow\begin{bmatrix}B & a & c & k & \\ t & o & & t & \overline{h}\\ e & & \overline{d} & r & a\\ w & \overline{i} & n & g & \\ b & o & a & r & \overline{d}\end{bmatrix}.$$

In view of the above, we note the derivation of a matrix-based data encryption method based on the Fibonacci polynomial matrix, in which, for the input message matrix over the alphabet {0, 1, …, 255} and for the integer $x = 2$, a third-order Fibonacci matrix is used, the elements of which are Fibonacci polynomials of degrees from 2 to 1. Leibniz's formula was used to calculate the determinant of the matrix, whilst the Montante method was applied to find the inverse matrix.

4. Conclusions

The possibility of solving the problem of generating the *n*th sequence of Fibonacci polynomials of degree *n*, which serve as the foundation for block data encryption enabling their effective protection, has been demonstrated. A method for generating a sequence of *n* modified Fibonacci polynomial matrices of order *m* has been developed, whose elements consist of Fibonacci polynomials with indices ranging from (*n*-2) to (2(*m*—2)+*n*). Unlike existing methods, this approach utilizes a matrix recurrent relation analogous to the recurrent relation used for generating Fibonacci numbers, while allowing both the determinants and inverse matrices of the corresponding order to be computed from the resulting matrices. A specific example is provided to illustrate the application of the matrix method for block data encryption using the 3rd Fibonacci polynomial matrix of order 5.

References

[1] Abd-Elhameed, Waleed Mohamed, Philippou, Andreas N., & Zeyada, Nasr Anwer. (2022). Novel Results for Two Generalized Classes of Fibonacci and Lucas Polynomials and Their Uses in the Reduction of Some Radicals. Mathematics, 10(11), article ID 2342. https://doi.org/10.3390/math10132342

[2] Esmaili, Mostafa, & Esmaeili, Morteza. (2010). A Fibonacci-polynomial based coding method with error detection and correction. *Computers and Mathematics with Applications*, Vol. 60, issue 10, 2738–2752. https://doi.org/10.1016/j.camwa.2010.08.091

[3] Falcon, S., & Plaza, A. (2009, February). On *k*-Fibonacci sequences and polynomials and their derivatives. *Chaos, Solitons and Fractals*, Vol. 39, issue 3, 1005–1019. https://doi.org/10.1016/j.chaos.2007.03.007

[4] Fox, William P., & West, Richard D. (2024, September). Numerical Methods and Analysis with Mathematical Modelling (Textbooks in Mathematics). Chapman and Hall/CRC, 424 p. URL: https://www.amazon.com/Numerical-Mathematical-Modelling-Textbooks-Mathematics/dp/1032697237

[5] Grytsiuk, P. Y., & Hrytsiuk, Y. I. (2025). Method for generating a sequence of Fibonacci polynomial matrices and their features for use in block data encryption. *Scientific Bulletin of UNFU*, 35(1), 173–191. https://doi.org/10.36930/40350121

[6] Hrytsiuk, Yu. I., & Grytsiuk, P. Yu. (2016). Features of generating Fibonacci Q_p-matrices – keys for implementing cryptographic transformations. *Bulletin of the National University "Lviv Polytechnic". Series: Computer Science and Information Technology*, Vol. 843, 251–263. URL: https://vlp.com.ua/node/16094

[7] Özkan, Engin, Taştan, Merve & Aydoğdu, Ali. (2019). Fibonacci Sayılarının Ailesinde 3-Fibonacci Polinomları. *Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi*. Cilt: 12 Sayı: 2, 926 –933. https://doi.org/10.18185/erzifbed.512100

[8] Stakhov, A. P. (2006, October). Fibonacci matrices, a generalization of the "Cassini formula", and a new coding theory. *Chaos, Solitons, & Fractals*, Vol. 30, issue 1, 56–66. https://doi.org/10.1016/j.chaos.2005.12.054

[9] Vajda, S. (2007, December). Fibonacci and Lucas Numbers, and the Golden Section. *Theory and Applications* (Dover Books on Mathematics). Dover Publications, 192 p. URL: https://www.amazon.com/Fibonacci-Lucas-Numbers-Golden-Section/dp/0486462765

[10] Yang, Jizhen, & Zhang, Zhizheng. (2018, December). Some identities of the generalized Fibonacci and Lucas sequences. *Applied Mathematics and Computation*, Vol. 339, 451–458. https://doi.org/10.1016/j.amc.2018.07.054

[11] Yilmaz, N. (2024). Some new results for the generalized bivariate Fibonacci and Lucas polynomials. Corpus ID: 271426809. Mathematica Moravica, Vol. 28, br. 1, 97–108. https://doi.org/10.5937/matmor2401097y

On Modern Theoretical Aspects of the Security of Syber-Physical Systems

Oksana Kholyavka[1,3], Nataliia Buhrii[2] and Oleh Buhrii[3]

[1] *Institute for Applied Problems of Mechanics and Mathematics, Naukova Str., 3b, 79060 Lviv, Ukraine*
[2] *Lviv Polytechnic National University, Mytropolyt Andrei Str., 5, 79000 Lviv, Ukraine*
[3] *Ivan Franko National University of Lviv, Universytetska 1, 79000 Lviv, Ukraine*

Abstract
We consider a mathematical model of the functioning of some cyber-physical systems. The model contains some parabolic partial differential equations. We provide a theoretical justification for an algorithm that can be used to detect cyberattacks on the system.

Keywords
cyber-physical system, inverse problem, parabolic equation, variable exponent of nonlinearity

1. Introduction

The security of Distributed Parameter Cyber-Physical Systems (DPCPS) is of great importance in the face of cyberattack threats. Mathematical modeling combined with advanced computer technologies allows us to detect cyberattacks and distinguish them from the internal changes in the dynamics of the considered systems. We investigate the security aspects of DPCPS from a theoretical perspective. In particular, we focus on DPCPS modeled using semilinear partial differential equations with the variable exponent of the nonlinearity.

Our results are strongly motivated by papers [1], where similar problems are considered for the linear equations (see also [2-5]).

2. Statement of problem

We provide a theoretical justification for an algorithm that can be used to detect cyberattacks on the Cyber-Physical Systems (CPS).

Let us consider CPS that contain a semilinear parabolic partial differential equations with the variable exponent of the nonlinearity. Equations with the variable nonlinearity indices are a powerful tool for studying phenomena of complex structure. They allow us to take into account the internal features of the system more accurately.

More precisely, let us consider an equation in two independent variables, one of which is a time variable and the other is a spatial variable. The equation is considered in a rectangle. Let us supplement this equation with Neumann-type boundary conditions for the spatial variable, an initial condition for the time variable, and an local overdetermination condition. We obtain a problem in which the unknowns are a pair of functions - the solution of the equation and the coefficient of the equation. In the theory of differential equations, such problems are called inverse problems.

Corresponding authors
oksana.kholiavka@lnu.edu.ua (O. Kholyavka); nataliia.v.buhrii@lpnu.ua (N. Buhrii); oleh.buhrii@lnu.edu.ua (O. Buhrii)
0000-0002-5843-7374 (O. Kholyavka); 0000-0002-1554-7303 (N. Buhrii); 0000-0002-1698-5559 (O. Buhrii)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In the work we carry out the following stages. First, we prove the unique solvability of the direct problem, assuming the coefficient of the equation to be known and rejecting the overdetermination condition. The direct problem is studied in a weak formulation, that is, the solution is obtained in generalized Lebesgue and Sobolev spaces. In particular, to prove the existence of a solution to the direct problem, the Faedo-Galerkin method is used. Various problem for evolution partial differential equations with the variable exponents of the nonlinearity are investigated in our previous papers [6-8] and in works [9-12].

Then, similarly to [1], we proceed to the study of the inverse problem. Having reduced it to the Volterra integral equation, we show the unique solvability of the problem. Further, we will show the continuous dependence of the solution of the inverse problem on the data from the initial condition.

Finally, we will use the obtained results to detect a cyberattack on an electric vehicle, or more precisely, on the battery of an electric vehicle.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Roy T., Dey S. "Security of distributed parameter cyber-physical systems: Cyber-attack detection in linear parabolic PDEs." *ArXiv Preprint ArXiv:2107.14159* (2021). https://arxiv.org/pdf/2107.14159

[2] Parashar M., Thorp J.S., and Seyler C.E. “Continuum modeling of electromechanical dynamics in large-scale power systems,” IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 51, no. 9, pp. 1848–1858, 2004. https://doi.org/10.1109/TCSI.2004.834480

[3] Ghanavati M., Chakravarthy A., and Menon P. P. Analysis of automotive cyber-attacks on highways using partial differential equation models. IEEE Transactions on Control of Network Systems. 2018. Vol. 5, N 4. pp. 1775-1786. https://doi.org/10.1109/TCNS.2017.2760865

[4] Reilly J., Martin S., Payer M., and Bayen A.M. Creating complex congestion patterns via multi-objective optimal free way traffic control with application to cyber-security. Transportation Research Part B: Methodological. 2016. Vol. 91, pp. 366-382. https://doi.org/10.1016/j.trb.2016.05.017

[5] Demetriou M.A. Detection of communication attacks on spatially distributed systems with multiple interconnected actuator/sensor pairs. In 2018 IEEE Conference on Decision and Control (CDC), pp. 2896-2901, IEEE, 2018. https://doi.org/10.1109/CDC.2018.8619173

[6] Buhrii O.M. Visco-plastic, Newtonian, and dilatant fluids: Stokes equations with variable exponent of nonlinearity. Matematychni Studii. 2018. Vol. 49, N 2. pp. 165-180. https://doi.org/10.15330/ms.49.2.165-180

[7] Buhrii O.M., Buhrii N.V., Kholyavka O.T. Hyperbolic Stokes system of the third order with variable exponent of nonlinearity. Journal of Mathematical Sciences. 2025. Vol. 291, N 5. pp 617-635. https://doi.org/10.1007/s10958-025-07841-x

[8] Buhrii O.M. Stochastic parabolic equations on graphs. Matematychni Studii. 2026. Vol. 65, N 1. pp. 58-73. https://doi.org/10.30970/ms.65.1.58-73

[9] Ruzicka M. Electrorheological fluids: modeling and mathematical theory. Springer, 2007. https://doi.org/10.1007/BFb0104029

[10] Antontsev S., Shmarev S. Evolution PDEs with nonstandard growth conditions. Existence, uniqueness, localization, blow-up. Atlantis Studies in Diff. Eq., Vol. 4, Paris: Atlantis Press, 2015. https://doi.org/10.2991/978-94-6239-112-3

[11] Radulescu V.D., Repovs D.D. Partial differential equations with variable exponents: variational methods and qualitative analysis. CRC press, 2015. https://doi.org/10.1201/b18601

[12] Kaltenbach A. Pseudo-monotone operator theory for unsteady problems with variable exponents. Lecture Notes in Mathematics. Vol. 2329. Springer Nature Switzerland AG 2023. https://doi.org/10.1007/978-3-031-29670-3

Adaptive Detection of Goal Conflicts in Multi-Agent Smart Home Cyber-Physical Systems

Ihor Krestianinov[1]

[1] *State University of Information and Communication Technologies, Kyiv, Ukraine*

Abstract

Modern smart homes are cyber-physical systems that integrate heterogeneous IoT devices, sensors, actuators, communication networks, and autonomous software agents. Their simultaneous operation may create not only functional and energy-management conflicts but also security-related conflicts. For example, an energy-management agent may attempt to switch off a device required by a security agent, while a comfort-oriented agent may request an action prohibited by an active security policy. Such interactions may result in unauthorized actions, security-policy violations, or unsafe physical states. This paper proposes an adaptive method for detecting goal conflicts in multi-agent smart home cyber-physical systems. The method combines semantic incompatibility of agent goals with shared-resource constraints, temporal dependencies, security policies, physical-state impact, and the current cyber-physical context. A security-aware conflict indicator is formulated to estimate conflict severity, while adaptive thresholds and predictive assessment are proposed for identifying potentially critical conflicts before corresponding actions are executed. The proposed method is intended as an additional detection layer for subsequent coordination, access-control, optimization, or predictive-control mechanisms..

Keywords

cyber-physical system, smart home, Internet of Things, multi-agent system, goal conflict, conflict detection, cybersecurity, adaptive detection, security policy

1. Security Challenges in Multi-Agent Smart Home Cyber-Physical Systems

Smart homes are increasingly implemented as interconnected cyber-physical environments in which IoT devices, sensors, actuators, communication networks, and software services interact to monitor and control physical processes. Their connectivity provides automation, energy management, comfort, and safety, but also increases cybersecurity and privacy complexity [1,2]. Unlike conventional information systems, smart-home cyber-physical systems combine computational decisions with physical processes, meaning that an incorrect or unauthorized software action can directly affect the physical environment [3].

The problem becomes more complex when smart-home components are represented by autonomous agents. A multi-agent smart home can contain specialized agents responsible for HVAC control, energy management, access control, surveillance, privacy, appliance scheduling, and user interaction. Each agent may have its own goals, constraints, and decision-making policy. Although these goals may be individually valid, their simultaneous execution can produce conflicts.

Goal conflict is a recognized problem in multi-agent systems. Barber et al. distinguish goal, plan, and belief conflicts and demonstrate the importance of identifying goal conflicts before plans are integrated [4]. Van Riemsdijk et al. provide formal foundations for representing conflicting goals through the semantics of desirable states [5]. However, these approaches do not explicitly consider the security consequences of goal conflicts in dynamic smart-home cyber-physical environments.

Corresponding author

i.krestiyaninov@stud.duikt.edu.ua (I. Krestianinov)

0009-0009-0338-0217 (I. Krestianinov)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In a smart home, a conflict can have direct security implications. An HVAC agent may attempt to open a window to improve thermal comfort while a security agent requires windows to remain closed because the building is in an armed state. An energy-management agent may attempt to disconnect a surveillance device to reduce electricity consumption, while a security policy requires that device to remain operational. A user agent may request access to a smart lock while the access-control agent must enforce a restrictive policy. Such conflicts do not necessarily result from malicious behavior; they may emerge from legitimate agents pursuing incompatible objectives.

Therefore, this work addresses the problem of designing an adaptive mechanism for detecting and evaluating conflicts between autonomous agent goals while taking into account their potential impact on security policies and physical states of a smart-home cyber-physical system.

2. Security-Aware Adaptive Conflict Detection

Existing multi-agent conflict-detection approaches primarily analyze incompatibilities between goals, plans, and beliefs [4,5]. Smart-home research has also investigated multi-objective optimization and adaptive management of energy consumption, comfort, and user preferences [6,7]. Recent studies consider multi-agent architectures and adaptive reasoning for heterogeneous smart-home IoT environments [7,8]. However, these approaches primarily focus on operational optimization and autonomous decision making rather than explicitly quantifying conflicts that may violate security policies or produce unsafe physical states.

The proposed approach represents goal conflict as a continuous, context-dependent, security-aware quantity. In addition to semantic goal incompatibility, it considers competition for shared cyber-physical resources, temporal dependencies, security-policy constraints, potential physical-state impact, and operational context. This representation is intended to distinguish between a theoretically incompatible pair of goals and a conflict that becomes critical under specific operating conditions.

The proposed method also introduces adaptive weighting and threshold selection. The importance of a particular conflict may change depending on occupancy, security mode, device state, energy conditions, time, and user priorities. Therefore, the conflict assessment parameters are treated as adaptive variables rather than fixed constants.

A further component is predictive conflict detection. Instead of evaluating only the current state, the method is designed to estimate whether agent goals may become incompatible during a future control horizon. This makes it possible to identify potentially unsafe or unauthorized combinations of actions before their execution.

The approach is intended to complement existing IoT security mechanisms rather than replace them. Authentication, authorization, secure communication, device hardening, and intrusion detection remain necessary, while the proposed method addresses conflicts generated by legitimate autonomous decisions.

3. Conflict Assessment Model and Adaptive Detection Procedure

Let a smart-home cyber-physical system contain a set of autonomous agents $A = A_1, A_2, \ldots, A_m$.

Each agent is associated with active goals, available actions, constraints and applicable security policies. The system state is represented as

$$X(t) = X_c(t), X_p(t), X_{ctx}(t) \quad (1)$$

where $X_c(t)$ denotes the cyber state, $X_p(t)$ denotes the physical state, and $X_{ctx}(t)$ denotes contextual information such as occupancy, time, energy demand, and user preferences.

For two agents A_i and A_j, the semantic incompatibility of their active goals is represented by the normalized goal-conflict indicator:

$$C_{goal}^{ij}(t) = D(G_i(t), G_j(t)), 0 \le C_{goal}^{ij}(t) \le 1 \quad (2)$$

where D is a compatibility function. A value close to zero indicates that the considered goals are compatible, whereas a value close to one indicates a high degree of semantic incompatibility.

The proposed method extends this basic measure by incorporating additional sources of conflict arising from the interaction of autonomous agents in the cyber-physical environment. The overall conflict score is defined as:

$$C_F^{ij}(t) = \sum_{k=1}^{6} w_k(t) C_k^{ij}(t) \quad (3)$$

where $C_F^{ij}(t)$ represents the corresponding conflict component. In the proposed model, these components describe goal incompatibility, shared-resource competition, temporal incompatibility, security-policy conflict, physical-state impact, and contextual influence.

To maintain a normalized conflict score, the weighting coefficients satisfy:

$$\sum_{k=1}^{6} w_k(t) = 1, w_k(t) \ge 0 \quad (4)$$

The security component is particularly important for the considered cyber-physical environment. It evaluates whether the simultaneous pursuit of two agent objectives may violate an active security policy, create an access-control conflict, or affect a security-critical IoT device. It can be represented as:

$$C_{sec}^{ij}(t) = alpha_1 V_{pol}^{ij}(t) + alpha_2 R_{acc}^{ij}(t) + alpha_3 R_{dev}^{ij}(t) \quad (5)$$

Where $V_{pol}^{ij}(t)$ represents the potential violation of an active security policy, $R_{acc}^{ij}(t)$ represents access-control risk, and $R_{dev}^{ij}(t)$represents the risk associated with a security-critical IoT device. The coefficients determine the relative contribution of these factors and can be normalized so that their sum equals one.

The physical-state component evaluates whether the combined actions of two agents may lead to an undesirable physical condition. For example, switching off a surveillance device may be beneficial from an energy-management perspective but incompatible with an active security requirement. This allows the proposed approach to account for the connection between software-level decisions and physical consequences.

The adaptive mechanism is intended to modify the relative importance of conflict components according to the current system context and feedback from previous decisions. A normalized update of the weighting coefficients can be expressed as:

$$w_k(t+1)=\frac{w_k(t)+\eta e_k(t)}{\sum_{r=1}^{6}[w_r(t)+\eta e_r(t)]} \tag{6}$$

where η denotes the adaptation rate and $e_k(t)$ represents feedback associated with the corresponding conflict component. This formulation provides a mathematical representation of the adaptive principle; the selection of η and the definition of the feedback signal require future experimental investigation.

The proposed method also includes predictive conflict assessment. Instead of evaluating only the current system state, the method is intended to estimate the conflict level associated with a predicted future state:

$$\hat{C}_F^{ij}(t+k)=F(G_i,G_j,P_i,P_j,X(t),\hat{X}(t+k)) \tag{7}$$

where $\hat{X}(t+k)$ denotes the predicted cyber-physical state at a future time $t+k$. The predicted conflict score can be used to identify potentially problematic interactions before the corresponding agent actions are executed.

The conflict score can then be interpreted using two adaptive thresholds:

$$L^{ij}(t)=\begin{cases}\text{compatible} & C_F^{ij}(t)<\theta_1(t)\\ \text{potential conflict} & \theta_1(t)\le C_F^{ij}(t)<\theta_2(t)\\ \text{critical conflict} & C_F^{ij}(t)\ge\theta_2(t)\end{cases} \tag{8}$$

where $\theta_1(t)$ and $\theta_2(t)$ are adaptive decision thresholds. The classification defines the intended operating principle of the proposed detector and does not represent experimentally validated performance levels.

These categories define the intended operating principle of the detector rather than experimentally validated performance levels. In the absence of experimental data, the threshold values cannot yet be associated with specific accuracy or security-performance guarantees.

The resulting conceptual architecture consists of agent goal analysis, contextual and security-policy assessment, physical-state assessment, adaptive conflict scoring, predictive classification, and subsequent coordination or security enforcement. The detector is intended to operate between the agent decision and action-execution layers. Depending on the detected conflict level, an action may proceed, require coordination, or be blocked or modified according to the applicable security policy.

The proposed method can potentially be integrated with multi-objective optimization, model predictive control, and reinforcement-learning approaches. In such an architecture, an optimization component can search for compromises between operational objectives, while the conflict detector can provide an additional security-related constraint. Experimental verification of these integrations is outside the scope of the present work.

4. Discussion and Conclusions

This paper has formulated an adaptive method for detecting goal conflicts in multi-agent smart-home cyber-physical systems. The proposed approach extends conventional goal-conflict analysis

by incorporating security-policy constraints and potential physical consequences of autonomous agent actions.

The central element of the method is a security-aware conflict score that combines semantic goal incompatibility, resource competition, temporal dependencies, security-policy conflicts, physical-state impact, and contextual information. The formulation allows the severity of a conflict to be represented as a continuous value rather than only as a binary compatible/incompatible relationship.

The method further proposes adaptive weights and thresholds whose values can be adjusted according to the current cyber-physical context and feedback from previous decisions. Predictive assessment is included to support the identification of potentially critical conflicts before the associated actions are executed.

At the current stage, the proposed method represents a conceptual and mathematical framework. No experimental evaluation has yet been performed; therefore, conclusions concerning detection accuracy, computational efficiency, false-positive or false-negative rates, or practical security improvement cannot be made. These characteristics require validation using a representative smart-home cyber-physical test environment.

Future research will focus on implementing the proposed detector and evaluating it using representative scenarios involving energy management, access control, surveillance, HVAC control, and security-critical IoT devices. The planned evaluation should investigate detection accuracy, detection lead time, computational overhead, and the ability to identify conflicts that may lead to security-policy violations or undesirable physical states.

Declaration on Generative AI

Either:
The author(s) have not employed any Generative AI tools.

Or (by using the activity taxonomy in ceur-ws.org/genai-tax.html):
During the preparation of this work, the author(s) used X-GPT-4 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Alsufyani, O. Rana, and C. Perera, "Knowledge-based cyber physical security at smart home: A review," ACM Computing Surveys, vol. 57, no. 3, article 53, 2024. doi:10.1145/3698768..

[2] D. Radovanovic, N. Savanovic, J. Janackovic, and P. Kresoja, "A Systematic Review of Smart Home IoT Security: Applications, Threat Taxonomy, Privacy Risks, and Emerging Defensive Solutions," Big Data and Cognitive Computing, vol. 10, no. 8, article 252, 2026. doi:10.3390/bdcc10080252.

[3] "Cyber-physical systems security: A systematic review," Computers & Industrial Engineering, vol. 188, article 109891, 2024. doi:10.1016/j.cie.2024.109891.

[4] K. S. Barber, T. H. Liu, and S. Ramaswamy, "Conflict detection during plan integration for multi-agent systems," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 31, no. 4, pp. 616–628, 2001. doi:10.1109/3477.938264.

[5] M. B. van Riemsdijk, M. Dastani, and J.-J. Ch. Meyer, “Goals in conflict: Semantic foundations of goals in agent programming,” Autonomous Agents and Multi-Agent Systems, vol. 18, pp. 471–500, 2009. doi:10.1007/s10458-008-9067-4.

[6] A. Anvari-Moghaddam, H. Monsef, and A. Rahimi-Kian, “Optimal smart home energy management considering energy saving and a comfortable lifestyle,” IEEE Transactions on Smart Grid, vol. 6, no. 1, pp. 324–332, 2015. doi:10.1109/TSG.2014.2349352.

[7] E. Dolinin and C. Kiourt, “Multi-Agent-Based Smart-Home Energy Management with Adaptive Reasoning,” Applied Sciences, vol. 16, no. 4, article 1896, 2026. doi:10.3390/app16041896.

[8] X. Du, F. Luo, J. Hu, W. Zhou, and J. Wen, “Adaptive home energy management to self-motivated user preferences via iterative LLM-augmented reinforcement learning,” Applied Energy, vol. 415, article 127976, 2026. doi:10.1016/j.apenergy.2026.127976.

Simulation of Fault Injection Attacks on a Secure Bootloader for the RISC-V Soft-IP Core (MIV_RV32) in Renode Framework

Liliia Melnychuk[1] and Mykola Shcherbyna[1]

[1] *Ivan Franko National University of Lviv, Universytetska 1, 79000 Lviv, Ukraine*

Abstract

Testing secure bootloaders against hardware fault injection attacks is notoriously non-deterministic and difficult to integrate into automated development pipelines. To address this challenge, we present software-based fault injection simulation as a scalable alternative to traditional laboratory equipment. Using the Renode emulator of a RISC-V platform, a methodology is proposed to deterministically induce instruction skips and identify vulnerable control-flow patterns, such as bypassed signature verification branches. This approach enables efficient, repeatable security validation and early-stage vulnerability assessment during secure firmware development.

Keywords

CI/CD, control-flow integrity, emulation, fault injection, hardware attack, MIV_RV32, Renode framework, RISC-V, secure bootloader, signature verification, simulation

1. Introduction

Testing embedded firmware, particularly secure bootloaders, against hardware fault injection attacks presents a significant challenge in modern cybersecurity. Secure bootloaders serve as the root of trust for embedded devices, enforcing cryptographic signature verification to ensure that only authentic and unmodified firmware is executed [1]. However, fault injection techniques can physically perturb the microcontroller during runtime, inducing transient hardware faults that alter the intended control flow. Such perturbations can cause the processor to skip critical instructions or corrupt register values during signature validation, effectively bypassing the security mechanisms [2]. Assessing the resilience against these sophisticated vectors is a critical, yet complex, component of the secure firmware development.

Clock glitching attack introduces rapid, anomalous disruptions into the target microcontroller's clock signal. By momentarily violating the operational timing constraints of the internal digital logic, an attacker can cause the CPU to misinterpret instructions, skip execution cycles, or incorrectly latch data [2]. Hardware evaluation of clock glitching vulnerabilities is typically conducted using specialized diagnostic equipment, such as the ChipWhisperer® Husky [3], which allows researchers to inject sub-nanosecond clock faults precisely synchronized with the target's execution flow.

Similarly, voltage glitching targets the power supply network of the integrated circuit. By rapidly pulling down the core operating voltage for a brief duration, the attacker starves the internal logic gates of adequate power exactly when a critical operation – such as a cryptographic signature verification – is occurring. This transient undervoltage induces setup and hold time violations, leading to computational errors or instruction corruption without triggering standard brownout reset protections [2]. As with clock-based attacks, platforms like the ChipWhisperer®

Corresponding authors

liliia.melnychuk@lnu.edu.ua (L. Melnychuk); mykola.shcherbyna@lnu.edu.ua (M. Shcherbyna)

0009-0008-9840-6768 (L. Melnychuk); 0009-0003-9761-9466 (M. Shcherbyna)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Husky are frequently employed to generate and synchronize these precise voltage transients against the device under test [3].

Electromagnetic Fault Injection (EMFI) represents a more localized, non-contact approach. Instead of manipulating global clock or power lines, EMFI utilizes a high-voltage pulse generator and a specialized coil probe to emit a concentrated electromagnetic field over a specific region of the semiconductor die. This intense, localized electromagnetic pulse induces eddy currents within the silicon, causing targeted bit flips or instruction corruption in specific internal hardware modules [2]. Tools such as the ChipSHOUTER CW520 (Figure 1) are standard in the industry for performing these spatially and temporally precise electromagnetic perturbations [4].



Figure 1: ChipSHOUTER CW520 being held by one of the authors. [CC BY 4.0], via the authors.

Despite the proven effectiveness of hardware-based fault injection, reproducing these attacks in a laboratory environment is inherently non-deterministic and highly sensitive to external variables. Successful physical fault induction often requires tens of thousands of iterative attempts to perfectly align the glitch parameters – timing, duration, and intensity – with the vulnerable instruction cycle. This instability makes hardware testing notoriously difficult to scale and integrate into automated Continuous Integration and Continuous Deployment (CI/CD) pipelines. To address this challenge, this paper proposes a software-based approach to simulate fault injection using the Renode framework [5]. By programmatically inducing instruction skips in a deterministic virtualized environment, we achieve stable, reproducible results, enabling automated and scalable vulnerability assessment of secure bootloaders during the software development phase.

2. Simulation of Fault Injection Attacks on MIV_RV32

2.1. RISC-V architecture and MIV_RV32 soft-core processor

RISC-V originated as an academic project at the UC Berkeley Parallel Computing Laboratory (Par Lab) led by Krste Asanović and David Patterson. The initial technical reports and specifications were published between 2011 and 2014. In 2015, governance of the standard was transferred to the RISC-V Foundation (currently RISC-V International). The core concept is an open, modular, royalty-free Instruction Set Architecture (ISA) that allows developers to freely implement processor cores and extend the base ISA with custom instructions.

RISC-V defines base profiles alongside numerous standard extensions. The base instruction sets comprise RV32I, RV64I, and RV128I, while extensions include M (integer multiplication and division),

A (atomic operations), F/D/Q (single, double, and quad-precision floating-point), C (compressed instructions), B (bit manipulation), V (vector operations), and other specialized "Z" extensions.

RV32I is the base 32-bit integer instruction set, designed to be minimal yet sufficient for target compilation and operating system support. Its architectural model features 32 general-purpose registers (x0–x31, where x0 is hardwired to zero), a program counter (PC), and a fixed 32-bit instruction encoding format, comprising approximately 40 core instructions [6].

MIV_RV32 is a commercial soft-core processor IP core from Microchip Technology designed for FPGAs. It is implemented as a configurable RV32I-compliant core with optional M (multiplication/division), C (compressed instructions), and F (single-precision floating-point) extensions depending on the core configuration [7].

2.2. Simulated Secure Bootloader

For demonstration purposes, a virtual system based on the Microchip Technology MIV_RV32 processor core (RISC-V architecture) was selected. Firmware samples were compiled using the GCC cross-compiler (riscv-none-elf-gcc, version 15.2.0). Since compilation targets bare-metal without a Real-Time Operating System (RTOS), Hardware Abstraction Layer (HAL) drivers from miv-rv32-bootloader were utilized.

Communication with the host system is established via a UART interface (115,200 baud, 8 data bits, 1 stop bit), initialized platform-dependently at the beginning of `main()`. Furthermore, to satisfy low-level C library porting requirements, the `_write(int fd, const char *buf, size_t count)` system call handler was overridden to transmit buffer content over UART.

For this evaluation, a simplified Secure Bootloader emulator was created. In this setup, the target firmware payload consists of 512 bytes populated with the repeated string "`FIRMWARE`". The simulated bootloader verifies payload integrity and authenticity using a standard cryptographic scheme: it first computes a SHA-256 digest of the firmware image, then verifies this digest using ECDSA over the NIST P-256 elliptic curve. In code, this logic is encapsulated by the `p256_verify` function call, which accepts the firmware binary, payload length, a signature in $r\|s$ concatenated format, and a public key represented as an uncompressed EC point. For this purpose, a lightweight third-party library [8] was integrated.

A private key on the curve was generated, and the corresponding public key was exported. In the output of `openssl ec -pubout -text -noout`, OpenSSL displays the public key as an uncompressed point (a `04:`... byte string followed by 32 bytes for the X-coordinate and 32 bytes for the Y-coordinate). These values were embedded into the `pub[]` array within the source code.

The `fake_fw.bin` file (corresponding to the `firmware[]` array) was signed using the command `openssl dgst -sha256 -sign ecdsa_priv.pem -out fake_fw.sig fake_fw.bin`. Subsequently, `openssl asn1parse -in fake_fw.sig -inform DER -i -dump` was executed to extract the hexadecimal values of r and s, which were inserted into the `sig[]` array.

The `p256_verify` execution flow performs the following actions:

- accepts the raw firmware byte array and payload length, computing the SHA-256 digest internally within the library;
- interprets the pub array as an elliptic curve point (X,Y) and reconstructs the public key;
- interprets the sig array as concatenated 32-byte r and 32-byte s values;
- utilizes the reconstructed public key and computed digest to verify signature validity against the signing key;
- returns `P256_SUCCESS` upon successful verification.

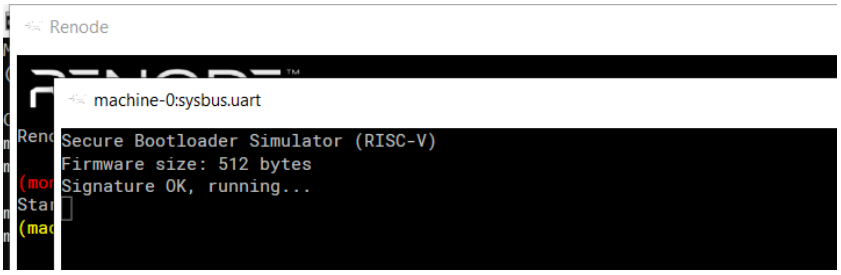


Figure 2: Successful firmware signature verification. [CC BY 4.0], via the authors.

In the event of verification failure, the program enters an infinite loop (simulating a panic condition). Upon success, a confirmation message is emitted and normal execution proceeds.

2.3. Fault injection attack simulation

The firmware binary `test_risc-v.bin` was patched by replacing the string segment "`FIRM`" with "`HACK`". To verify the modification, the script was launched in Renode, confirming that signature verification subsequently failed as expected.

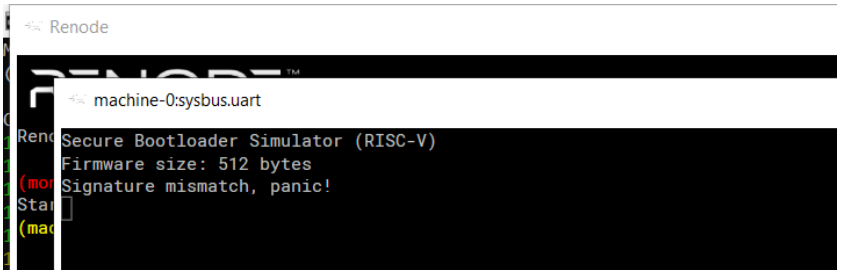


Figure 3: Failed firmware signature verification. [CC BY 4.0], via the authors.

To evaluate system resilience, a hardware fault injection attack was simulated to induce the bootloader into treating the tampered firmware payload as authentic. A Python script, fuzzing.py, was developed to automate software-based fault injection simulation by skipping instruction execution and identifying vulnerable instruction offsets within the bootloader routine. The script sequentially iterates through target addresses within a defined memory window based on command-line arguments:

- start address (`--begin`);
- end address (`--end`);
- platform instruction alignment (`--step`), corresponding to instruction width (4 bytes for 32-bit RISC-V);
- emulator script path (`--resc`);
- log file path (`--logfile`);
- per-iteration execution timeout (`--timeout`).

The primary mechanism for fault simulation involves attaching a dynamic hook to the target instruction address using the built-in `cpu AddHook` command. When the processor Program Counter (PC) hits this address, the script executes the monitor command `monitor.Parse('sysbus.cpu PC <new_pc>')`. This command overwrites the `PC` register

with the subsequent instruction address (`addr + step`), effectively skipping target instruction execution. If the bootloader bypasses signature verification following this artificial fault and outputs the target string "`Signature OK, running`" to the log, the script records the offset as vulnerable to fault injection attacks.

The Renode script `fuzz_risc-v.resc` redirects UART output to a file with unbuffered flushing enabled to ensure immediate log updates for post-execution analysis. The simulation was executed via:

```
python fuzzing.py --resc fuzz_risc-v.resc --logfile fuzz_risc-v.log --begin 0x600009a4 --end 0x60000a00 --step 4
```

producing the following output:

```
[+] Testing addr 0x600009a4 (skipping to 0x600009a8)... -> No success
[+] Testing addr 0x600009a8 (skipping to 0x600009ac)... -> No success
[+] Testing addr 0x600009ac (skipping to 0x600009b0)... -> No success
[+] Testing addr 0x600009b0 (skipping to 0x600009b4)... -> No success
                                        …
[+] Testing addr 0x600009d4 (skipping to 0x600009d8)... -> No success
[+] Testing addr 0x600009d8 (skipping to 0x600009dc)... -> No success
[+] Testing addr 0x600009dc (skipping to 0x600009e0)... -> SUCCESS: Signature OK
[+] Testing addr 0x600009e0 (skipping to 0x600009e4)... -> No success
[+] Testing addr 0x600009e4 (skipping to 0x600009e8)... -> No success
                                        …
[+] Testing addr 0x600009fc (skipping to 0x60000a00)... -> No success
[+] Testing addr 0x60000a00 (skipping to 0x60000a04)... -> No success

Summary: addresses with 'Signature OK, running'
0x600009dc
```

2.4. Disassembly analysis

Let us analyze the compiler-generated assembly code:

```
600009a4:   60018537        lui     a0,0x60018
600009a8:   ff010113        addi    sp,sp,-16
600009ac:   86818593        addi    a1,gp,-1944 # 80000068 <plat.0>
600009b0:   20000613        li      a2,512
600009b4:   6f050513        addi    a0,a0,1776 # 600186f0 <__clzsi2+0x8c>
600009b8:   00112623        sw      ra,12(sp)
600009bc:   219050ef        jal     600063d4 <printf>
600009c0:   600196b7        lui     a3,0x60019
600009c4:   ac468693        addi    a3,a3,-1340 # 60018ac4 <pub>
600009c8:   04468613        addi    a2,a3,68
600009cc:   08468513        addi    a0,a3,132
600009d0:   20000593        li      a1,512
600009d4:   3f4040ef        jal     60004dc8 <p256_verify>
600009d8:   00100793        li      a5,1
600009dc:   00f51c63        bne     a0,a5,600009f4 <test+0x50>
600009e0:   00c12083        lw      ra,12(sp)
600009e4:   60018537        lui     a0,0x60018
600009e8:   72c50513        addi    a0,a0,1836 # 6001872c <__clzsi2+0xc8>
600009ec:   01010113        addi    sp,sp,16
600009f0:   3990506f        j       60006588 <puts>
600009f4:   60018537        lui     a0,0x60018
600009f8:   74850513        addi    a0,a0,1864 # 60018748 <__clzsi2+0xe4>
600009fc:   38d050ef        jal     60006588 <puts>
60000a00:   fa1ff0ef        jal     600009a0 <panic>
```

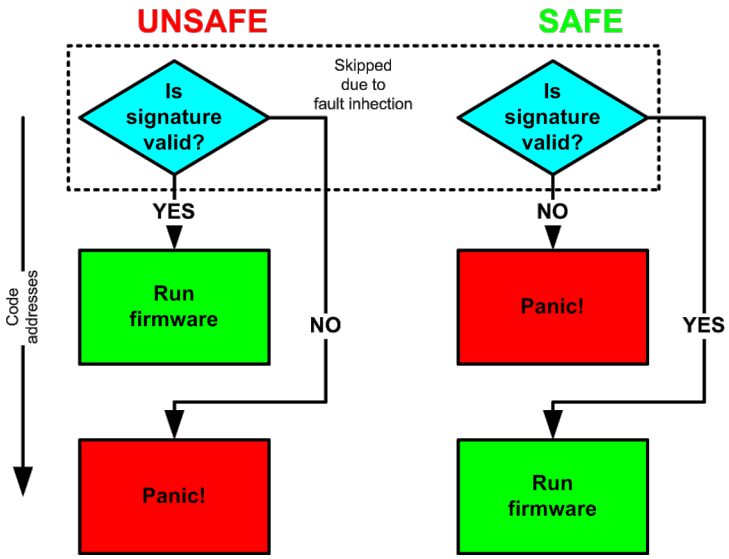


Figure 4: Unsafe and safe code patterns. [CC BY 4.0], via the authors.

At address `0x600009dc`, a branch-if-not-equal instruction (`bne`) is located. As a result of the fault injection, the processor skipped the execution of this instruction, allowing control flow to fall through and execution to proceed as if the signature verification had succeeded.

An interesting yet straightforward observation is that certain code patterns are more resistant to fault-injection attacks (Figure 4). In the safe scenario, if the comparison is skipped, the subsequent panic block is still executed, thereby mitigating the fault injection. We achieved this behavior by making minor edits to the C bootloader source. However, there is no deterministic way to force the compiler to consistently generate such fault-resistant machine code.

3. Conclusion

While contemporary security research focuses on physical fault injection attacks, systematic testing against such vulnerabilities remains underexplored. In this work, we successfully demonstrated a software-based fault injection testing framework utilizing the Renode emulator on the MIV_RV32 RISC-V platform. The proposed methodology can be expanded to bypass several consecutive instructions, deterministically simulate register corruption, and parallelize fault injection simulations, enabling automated validation within CI/CD pipelines.

Our findings indicate that minor source code modifications and compiler version variations can yield unexpected alterations in the generated binary code layout, shifting critical instruction boundaries. An alternative approach to physical hardware testing and our emulation framework is the automated verification of disassembly listings to identify vulnerable control-flow patterns. Although manual inspection is impractical at scale, this task is well-suited for automation using modern artificial intelligence (LLM) tools, presenting a promising avenue for future research.

Declaration on Generative AI

During the preparation of this work, the authors used Google Gemini in order to translate text and improve writing style. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Yu. Shcherbyna, P. S. Venherskyi, Vulnerabilities of the initial bootloader in microcontrollers with SPI flash memory, in: Proceedings of the XIV International Scientific and Technical Conference "ITSec: Bezpeka Informatsiinykh Tekhnolohiy", Ternopil, Ukraine, May 22–24, 2025, ZUNU-DUIKT, Ternopil-Kyiv, 2025, pp. 220–222. URL: https://drive.google.com/file/d/1Nt2w9E-N97TAIdAGqWKbaXsqASyP-4tt/view?usp=sharing.

[2] J. van Woudenberg, C. O'Flynn, The Hardware Hacking Handbook: Breaking Embedded Security with Hardware Attacks, No Starch Press, 2021.

[3] NewAE Technology, ChipWhisperer-Husky, 2026. URL: https://chipwhisperer.readthedocs.io/en/latest/Capture/ChipWhisperer-Husky.html.

[4] NewAE Technology, ChipSHOUTER, 2026. URL: https://chipwhisperer.readthedocs.io/en/latest/Starter%20Kits/ChipSHOUTER.html.

[5] Antmicro, Renode framework repository, GitHub, 2026. URL: https://github.com/renode/renode.

[6] RISC-V International, RV32I Base Integer Instruction Set, Version 2.1, 2026 URL: https://docs.riscv.org/reference/isa/v20240411/unpriv/rv32.html.

[7] Microchip Technology, MIV_RV32 User Guide, 2023. URL: https://ww1.microchip.com/downloads/aemDocuments/documents/FPGA/ProductDocuments/SupportingCollateral/MIV_RV32+v3.1+User+Guide.pdf.

[8] O. Reparaz, p256: ECDSA P-256 signature verification in a single C file targeting embedded use, GitHub repository, 2026. URL: https://github.com/oreparaz/p256.

Integration of neural network detectors into visual odometry systems

Mykhailo Moroz[1]

[1] State University of Information and Communication Technologies, Solomyanska 7, 02000 Kyiv, Ukraine

Abstract

The problem of accurate positioning and spatial mapping for autonomous systems operating in fast and dynamic environments is still a challenging task in computer vision. Traditional simultaneous localization and mapping (SLAM) systems rely heavily on the assumption of a static environment, expecting that all extracted visual features belong to fixed structures. When moving objects enter the camera's field of view, this assumption is violated, leading to false feature matching, significant trajectory drift, and ultimately tracking failure, as the algorithm misinterprets the object's motion as the camera's own motion. Modern hybrid systems attempt to address this problem and add deep learning for semantic segmentation. However, such systems often rigorously filter out features. This approach removes all properties that are important for dynamic objects. As a result, the system permanently loses data about those objects that have the meaning of being dynamic but are not moving at the current time. This could be, for example, parked cars in a city parking lot or people waiting at a crosswalk. Such information hunger inevitably leads to ill-conditioned matrices during optimization. This paper aims to solve this fundamental problem.

Keywords

SLAM; semantic segmentation; convolutional neural networks; covariance adjustment

1. Introduction

Simultaneous Localization and Mapping (SLAM) is essential for the autonomous operation of robotic platforms in GPS-denied environments. Classical visual odometry algorithms are highly accurate but rely on a strict assumption: the environment must remain entirely static. When moving entities enter the camera's field of view, they generate an optical flow distinct from the camera's ego-motion. Optimizing the platform's pose using these displaced features leads to severe miscalculations, rapid trajectory degradation, and ultimate localization failure.

To address this, modern systems integrate deep learning to semantically classify and exclude dynamic objects from the optimization graph. However, the interaction between neural network outputs and probabilistic SLAM algorithms, like the Extended Kalman Filter (EKF), remains suboptimal. Most approaches rely on rigid binary rejection: if an object belongs to a potentially dynamic class (e.g., a vehicle), all associated spatial features are unconditionally discarded.

The fundamental flaw is that convolutional neural networks recognize semantic classes without evaluating the actual kinematic state. In scenarios like navigating near a parking lot or a traffic red light, perfectly stationary vehicles are deleted despite being excellent geometric landmarks. This rigid filtering induces critical information starvation, mathematically crippling the algorithm's ability to compute the correct camera pose. Developing a methodology that preserves the speed of single-stage detectors while intelligently retaining static features on potentially dynamic objects remains a critical unresolved challenge.

Corresponding author

m.moroz@stud.duikt.edu.ua (M. Moroz)

0009-0002-9546-394X (M. Moroz)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Statement of the problem

The core problem addressed in this research is the suboptimal interaction between deep learning-based semantic segmentation and foundational probabilistic SLAM algorithms, specifically the Extended Kalman Filter (EKF) (2). Current data fusion mechanisms contain fundamental conceptual flaws that lead to the irreversible loss of critical spatial information (3). Resolving this requires a new algorithmic approach to integrate these systems without discarding valuable geometric data in complex environments (1). By developing a probabilistic data fusion mechanism that intelligently penalizes only genuinely moving objects, the system can preserve critical static reference points located on potentially dynamic targets, such as parked vehicles or stationary pedestrians. Achieving this continuous, adaptive weighting prevents the information starvation and matrix ill-conditioning that typically plague optimization in these scenarios. Ultimately, solving this problem ensures the navigational resilience and tracking continuity of autonomous unmanned systems—such as UAVs and ground robots—allowing them to operate without critical failures in highly dynamic environments, including disaster zones, search and rescue operations, and complex industrial inspections where satellite navigation is unavailable.

3. Solution to the problem at hand

To overcome the established limitations of binary feature rejection, this research proposes a probabilistic data fusion algorithm integrated directly into the update step of the Extended Kalman Filter (EKF). The core mechanism relies on dynamically adjusting the system's trust in specific visual features by combining the semantic confidence index generated by a single-stage YOLO detector—which ranges from 0 to 1 based on class certainty—with real-time kinematic data. In standard EKF implementations, an algorithmic arbiter known as the Kalman gain dictates how heavily a new visual observation should influence the correction of the robot's predicted trajectory. This gain is calculated by weighing the uncertainty of the system's internal kinematic model against the measurement noise of the incoming visual data. Traditionally, this visual measurement noise is treated as a static baseline value applied uniformly across the scene (4).

Our approach departs from this convention by dynamically modifying the measurement noise covariance for each individual visual feature through a nonlinear exponential penalty function. Rather than outright deleting a feature associated with a potentially dynamic class (such as a vehicle or a person), the algorithm inflates its assumed measurement noise based on its actual physical movement. This movement is quantified by a motion indicator, which is calculated as the average magnitude of the optical flow for the points located on that specific object.

This dynamic penalty is governed by three primary components: a baseline error representing the inherent hardware noise of the camera sensor, a regularization coefficient that dictates how rapidly the exponential penalty scales, and a global scaling coefficient that establishes the upper limit of the artificially induced noise. From a mathematical perspective, determining the optimal values for these hyperparameters relies on nonlinear programming. This critical offline calibration step is performed using Bayesian optimization or grid search on reference datasets with ground-truth trajectories, where the objective function is the global minimization of the Root Mean Square Error (RMSE).

Behaviorally, the system reacts adaptively to the actual kinematic state of the recognized object. If the object is actively moving, the motion indicator triggers the exponential penalty to spike, drastically inflating the object's assumed measurement noise. Consequently, the Kalman gain for these specific displaced features approaches zero, and the system dynamically ignores them during optimization process.

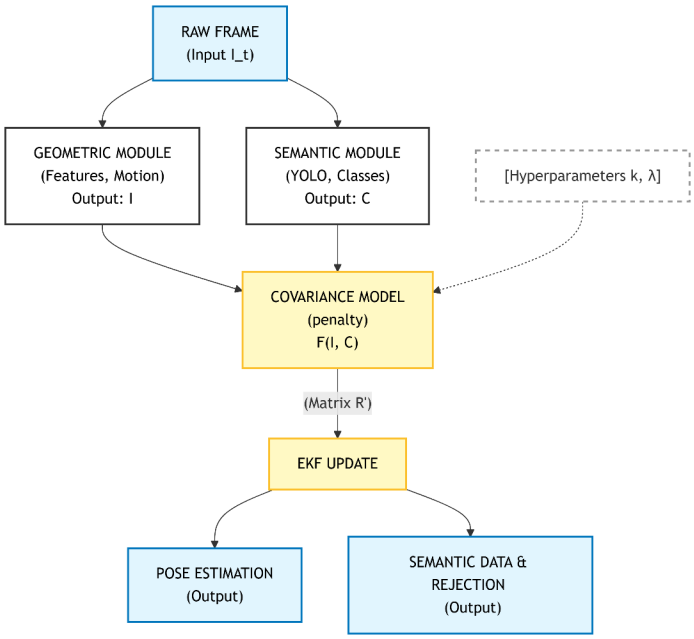


Figure 1: Schematic representation of the probabilistic data fusion method

Conversely, if the recognized object is physically stationary—despite belonging to a dynamic semantic class—the motion indicator and subsequent penalty multiplier are neutralized. The system then processes the feature at the baseline sensor noise level, seamlessly retaining it as a highly reliable static geometric landmark.

4. Evaluating obtained results

The proposed combined method was tested and compared against classical geometric approaches (ORB-SLAM3) (7) and binary-semantic rejection methods (DynaSLAM) (8). The primary comparison criterion was the trajectory Root Mean Square Error (RMSE), obtained through simulations using the public EuRoC dataset.

Table 1
Comparative RMSE of Methods (meters)

Dataset	Proposed Method	DynaSLAM	ORB-SLAM3
MH_01_easy	0.05	0.08	0.016
MH_02_easy	0.04	0.07	0.15
MH_03_easy	0.05	0.06	0.14
MH_04_difficult	0.08	0.11	0.25
MH_05_difficult	0.10	0.16	0.32
MH_06_difficult	0.06	0.07	0.13

The error values of the proposed method are, on average, 5-15% lower on complex datasets compared to the alternative approaches (Table 1). The most significant improvements in accuracy are recorded in sequences characterized by highly dynamic environments and rapid movements.

In these specific scenarios, the classical ORB-SLAM3 algorithm demonstrates a critical accumulation of positional error due to the unfiltered influence of dynamic obstacles. Conversely, the DynaSLAM system's rigid rejection mechanism artificially creates a deficit of reference features by discarding points on objects that are semantically dynamic but physically stationary at the time of observation.

In contrast, the implementation of the soft penalty model guarantees the preservation of a robust set of reliable static landmarks. The continuous, smooth modification of the statistical weights of these observations ensures the stable functioning of the optimization algorithm and prevents abrupt, erratic deviations in the calculated trajectory.

5. Conclusions

Proposed visual SLAM optimization method utilizing high-speed single-stage neural networks (YOLO). By demonstrating that rigid feature rejection degrades tracking in specific scenarios, we introduced an adaptive weighting mathematical model that seamlessly fuses semantic prediction probabilities with geometric data directly within the Kalman filter. Comparative analysis confirms this approach reduces trajectory RMSE, outperforming both traditional geometric and binary-semantic analogs. The application of this method significantly enhances the reliability of autonomous robotic navigation

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 3.1 in order to: Generate markdown mermaid code to render image represented in Figure 1.

References

[1] Mur-Artal, R., Montiel, J. M. M., & Tardos, J. D. (2015). ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE Transactions on Robotics, 31(5), 1147–1163. https://doi.org/10.1109/tro.2015.2463671.

[2] Choi, K.-S., & Lee, S.-G. (2010). Enhanced SLAM for a mobile robot using extended Kalman Filter and neural networks. International Journal of Precision Engineering and Manufacturing, 11(2), 255–264. https://doi.org/10.1007/s12541-010-0029-9

[3] Bescos, B., Fácil, J. M., Civera, J., & Neira, J. (2018). DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes. IEEE Robotics and Automation Letters, 3(4), 4076–4083

[4] Wadud, R. A., & Sun, W. (2022). DyOb-SLAM: Dynamic Object Tracking SLAM System. arXiv. https://doi.org/10.48550/arXiv.2211.01941

[5] Jocher, G., et al. (2024). YOLO by Ultralytics (Version 11). https://github.com/ultralytics/ultralytics

[6] Wang, C., et al. (2022). Applying SLAM Algorithm Based on Nonlinear Optimized Monocular Vision and IMU in the Positioning Method of Power Inspection Robot in Complex Environment. Mathematical Problems in Engineering, 2022, 1–14. https://doi.org/10.1155/2022/3378163

[7] Kim, A., Osep, A., & Leal-Taixe, L. (2021). EagerMOT: 3D Multi-Object Tracking via Sensor Fusion. IEEE International Conference on Robotics and Automation (ICRA).

[8] Liu, H. (2023). Identifying and updating local optimization methods in extended Kalman filter SLAM. Applied and Computational Engineering, 4(1), 569–573. https://doi.org/10.54254/2755-2721/4/2023325

Modeling Cyber-Physical Consequences of Additive Attacks on the Control Channel of an Aerodynamic Levitation System

Oleksii Novikov[1], Mykola Ilin[1], Iryna Stopochkina[1] and Volodymyr Duduladenko[1]

[1] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Beresteyskyi Ave 37, 05036 Kyiv, Ukraine*

Abstract

This paper investigates the cyber-physical consequences of additive attacks on the control channel of an aerodynamic levitation system controlled by a PLC with a PID controller. In contrast to attacks on the measurement channel, the height-sensor readings remain undistorted, while the attacker modifies the command after it is generated by the controller and before it reaches the fan actuator. A nonlinear three-state model parameterized from experimental data is used to represent ball motion, actuator-channel inertia, nonlinear aerodynamic lift, and height-dependent lift effectiveness. Four attack profiles are compared: constant bias, exponential disturbance, pseudorandom binary sequence (PRBS), and harmonic disturbance with time-varying frequency. The results show that standard feedback substantially compensates slow additive influences, whereas rapid sign-changing disturbances interact with actuator inertia and saturation and can produce much larger physical deviations. The PRBS attack caused the largest maximum height deviation, 0.3267 m. The results support joint monitoring of PLC-command integrity, actuator state, model residuals, and physical constraints.

Keywords

industrial control systems, PLC, PID, aerodynamic levitation, control-channel cyberattack.

1. Problem Statement and Scientific Novelty

Industrial control systems form a technological foundation of critical infrastructure, where violations of information integrity can directly produce physical consequences. In a closed automatic-control loop, a cyber impact may change not only the information presented to an operator but also the actual behavior of the technological process. Therefore, security analysis should cover sensor channels, PLC logic, command transmission, actuators, and the physical dynamics of the plant [1–5].

The study uses the Smart Process Control Learning System (2 × Air Fan) aerodynamic levitation setup [8] (Fig.1). The ball moves in a vertical air channel, and its height is regulated by changing the fan output. Aerodynamic levitation is convenient for cyber-physical experiments because nonlinear process dynamics are clearly observable and an incorrect control action is directly manifested in the physical motion of the ball [9, 10].

The authors previously investigated a complementary threat model in which an attack modifies the measurement channel, causing the PLC to receive a corrupted height value and to generate control from an incorrect error signal [2]. Another previous study considered stealthy attacks on control systems based on adaptive soft-constrained optimization [7]. The scientific novelty of the present work is the analysis of a different attack location and interaction mechanism: the sensor measurement remains correct, the PLC computes the real control error, but the command is modified downstream of the controller before reaching the actuator. Consequently, the PID controller observes the physical deviation caused by the attack and attempts to compensate for it

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

through standard negative feedback. The resulting dynamics are determined by the interaction among the additive influence, the compensating PID response, actuator-channel inertia, saturation, and the nonlinear plant.



Figure 1: Smart Process Control Learning System (2 × Air Fan) [9]. Photo by the authors.

The aim is to model and comparatively analyze the cyber-physical consequences of structured additive attacks on the control channel of the levitation system and to identify diagnostic features that distinguish severe control-channel disturbances from influences that are largely suppressed by the closed loop.

2. Model and Control-Channel Attack

The model is based on the balance of aerodynamic lift, gravity, and drag and uses the state vector $x=[h, v, u_f]^T$, where h is ball height, v is vertical velocity, and u_f is the effective actuator-channel state. The nonlinear dynamics used in the simulations are

$$\dot{h}=v,\ \dot{v}=\frac{(k_0+k_1u_f+k_2u_f^2)\mathrm{e}^{-k_h h}-mg-bv-c_d v|v|}{m}\ \dot{u}_f=(u_{act}-u_f)/\tau_a.,$$

Model parameters were identified from synchronized experimental histories of the measured ball height and fan control signal as described in the authors' previous work [2].

The measured output is the actual ball height. In the attacked loop, the PLC therefore continues to compute the correct control error and generates its command $u_{PID,att}$ according to the standard PID law [11]. The attack signal u_a is inserted after this command is generated (Fig.2). The command that reaches the actuator channel is

$$u_{att}(t)=sat_{[0,1]}[u_{PID,att}(t)+u_a(t)],$$

where $sat_{[0,1]}(.)$ is saturation operator.

After the attack changes the physical trajectory, the PID controller changes its own command in the opposite direction and may partially compensate for the disturbance. This feedback reaction is intrinsic to the investigated threat model and differs fundamentally from the measurement-channel case [2].

The identified model parameters used in the simulations were m = 0.06000001 kg, b = 0.29000001 $N{\cdot}s/m$, c_d= 1.02000001 $N{\cdot}s^2/m^2$, k_0 = 0.5179681 N, k_1 = 0.1 N, k_2 = 0.6 N, τ_a = 0.234000013 s, g = 9.81 m/s^2, and k_h = 0.4 m^{-1}; the reference height varied from 0.50 to 1.40 m.

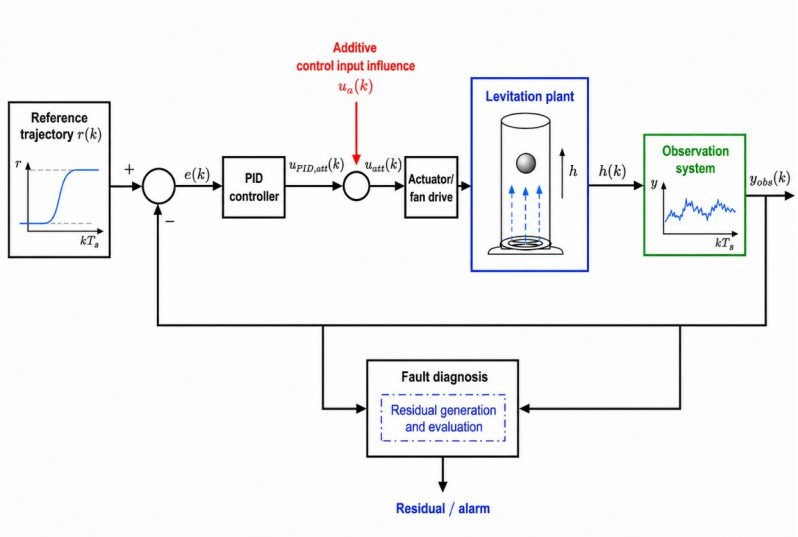


Figure 2: Control-system structure with an additive attack on the control channel.

The computational experiment used N = 573 samples over 36.6667 s with Δt = 0.0641 s. The PID gains were K_p = 6.0, K_i= 0.24, and K_d = 0.20; the normalized actuator command was constrained to [0, 1]. Four additive influences with the same nominal amplitude of 0.30 were investigated: a constant bias; an exponential disturbance with rise time τ_e = 5 s; a PRBS disturbance with 0.6 s switching interval and seed 12345; and a harmonic chirp whose frequency increased linearly from 0.2 to 2.0 Hz. The attack was active throughout the simulation. Cyber-physical impact was evaluated from the difference $\Delta h(t) = h_{att}(t) - h_{nom}(t)$, its maximum absolute value and RMS value, together with the fraction of time the actual actuator command remained near the lower or upper saturation limit.

3. Results and Implications for Secure Control

Table 1 summarizes the quantitative results. Because all scenarios use the same attack amplitude, the comparison isolates the influence of temporal structure and its interaction with the closed-loop dynamics. The A is attack amplitude.

Table 1
Quantitative indicators of the cyber-physical effect of the considered attacks.

Scenario / parameters	max|Δh|, m	RMS Δh, m	$u_{act}\approx 0$	$u_{act}\approx 1$
Constant bias, A = 0.30	0.0815	0.0282	0	0
Exponential, A = 0.30, τ = 5 s	0.0343	0.0245	0.17%	0
PRBS, A = 0.30, Δt_{sw} = 0.6 s	0.3267	0.1246	4.54%	21.82%
Chirp, A = 0.30, f0=0.2, f1=2.0 Hz	0.1048	0.0395	4.36%	31.24%

The constant positive bias increases the command supplied to the actuator, but the undistorted height measurement allows the PID controller to reduce its own command and counteract the attack.

The PRBS disturbance produces a qualitatively different response. Its repeated sign changes occur on a time scale comparable with the actuator and closed-loop dynamics: the controller reacts to the physical consequence of a previous switching event while the injected influence may already have changed sign. This interaction, together with actuator-channel inertia, produces the largest physical effect among the considered cases. The maximum deviation reaches 0.3267 m and the RMS deviation 0.1246 m; the actual command remains near the lower saturation limit for about 4.54% of the time and near the upper limit for about 21.82%.

The chirp attack successively excites the loop over an increasing frequency range. At low frequencies the PID controller has sufficient time to compensate, while the effect of actuator inertia and plant dynamics becomes more pronounced as frequency rises. The maximum height deviation is 0.1048 m. Notably, the actuator remains near the upper saturation limit for about 31.24% of the simulation, which is longer than in the PRBS case, although the physical deviation is much smaller. Therefore, saturation duration alone is not a sufficient measure of cyber-physical severity. The frequency content and temporal structure of the attack, actuator-channel inertia, controller response, and mechanical dynamics must be considered jointly.

These results have direct diagnostic implications. During a control-channel attack the measured height can remain reliable, while the command produced by the PLC differs from the action actually applied to the actuator. If both signals are available, the residual $r_u(t)=u_{att}(t)-u_{PID,att}(t)$ provides a direct indicator of command modification. If direct access to the actuator command is unavailable, model-based consistency monitoring can compare the measured physical response with the response predicted from the PLC command [5, 6, 12]. A significant model residual in the absence of a sensor anomaly indicates that the plant is receiving an input different from the one assumed by the controller.

Two complementary residuals follow directly from this interpretation. When the actual actuator command is observable, $r_u(t)=u_{att}(t)-u_{PID,att}(t)$ should remain close to zero in the nominal case and becomes sensitive to downstream command modification. When the actuator command is not directly accessible, a model or observer driven by the PLC command can generate an expected trajectory $\hat{h}_c(t)$, giving the physical residual $r_h(t)=h_{meas}(t)-\hat{h}_c(t)$. Neither residual should be interpreted in isolation, because similar deviations may also arise from actuator faults, communication disturbances, model uncertainty, or noise. Their diagnostic value increases when they are combined with sensor-consistency checks and the observed physical state of the process.

Accordingly, protection against this class of attacks should combine PLC-command integrity monitoring and authentication of control messages with feedback monitoring of the actual actuator state, model-residual analysis, and physical-constraint monitoring. Saturation can serve as an additional indicator of abnormal system behavior, but it should be interpreted together with residual signals and the observed physical deviation. The experiments also demonstrate that conventional feedback control can partially compensate for slow control-channel attacks. However, this compensating effect is limited and can be overcome by disturbances with an appropriate structure and magnitude.

4. Conclusions

This work analyzed additive attacks applied to the control channel of an aerodynamic levitation system after the PID command is generated. Unlike measurement-channel attacks previously

considered by the authors [2], the height signal remains undistorted; therefore, the PLC observes the true physical deviation and attempts to compensate for the attack. The analysis used an experimentally parameterized nonlinear three-state model that includes nonlinear lift, height dependence, actuator-channel inertia, and command saturation.

Four structured attack profiles were compared at the same amplitude of 0.30. Slow constant and exponential disturbances were largely suppressed by the closed loop, whereas the PRBS attack produced the largest height deviation, 0.3267 m. The chirp attack caused a smaller maximum deviation despite a larger fraction of operation near the upper saturation limit. Hence, cyber-physical impact cannot be assessed from attack amplitude or saturation time alone; the temporal and frequency structure of the injected signal and the dynamic properties of the controller, actuator, and plant are decisive. For detection, the most informative strategy is joint monitoring of PLC-command integrity, actual actuator behavior, model-based residuals, and physical constraints. Future work may compare the sensitivity of measurement and actuator channels to structured cyber-physical influences and relate these findings to more adaptive attack-generation mechanisms such as those studied in [7].

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI GPT-5.6 solely to format the manuscript in accordance with the conference template. The scientific content, research results, analysis, and conclusions were produced by the authors and were reviewed and verified by them. The AI tool was not used to generate or validate the research results.

References

[1] Sanchez H. S., Rotondo D., Escobet T., Puig V., Quevedo J. Bibliographical Review on Cyber Attacks from a Control Oriented Perspective. Annual Reviews in Control. 2019. Vol. 48. P. 103 –128. DOI: 10.1016/j.arcontrol.2019.08.002.

[2] Novikov O.M., Ilin M.I., Stopochkina I.V., Duduladenko V.M. Modeling Cyber-Physical Consequences of Attacks on the Measurement Channel of an Aerodynamic Levitation System. Theoretical and Applied Cyber Security. 2026. Vol. 8, No. 2.

[3] Wang Z., Humayed A., Li F., Luo B. A Survey on Programmable Logic Controller Vulnerabilities, Attacks, Detections, and Forensics. Processes. 2023. Vol. 11, No. 3. Article 918. DOI: 10.3390/pr11030918.

[4] O. Novikov, M. Shreider, I. Stopochkina, M. Ilin, Cyber Attacks Simulation for Modern Energy Facilities, in: Selected Papers of the XXIII International Scientific and Practical Conference “Information Technologies and Security” (ITS 2023), CEUR Workshop Proceedings, volume 3887, 2023, pp. 35–49. URL: https://ceur-ws.org/Vol-3887/paper4.pdf.

[5] Giraldo J., Urbina D., Cárdenas A. et al. A Survey of Physics-Based Attack Detection in Cyber-Physical Systems. ACM Computing Surveys. 2018. Vol. 51, No. 4. Article 76. P. 1–36. DOI: 10.1145/3203245.

[6] Leite D., Andrade E., Rativa D., Maciel A.M.A. Fault Detection and Diagnosis in Industry 4.0: A Review on Challenges and Opportunities. Sensors. 2025. Vol. 25, No. 1. Article 60. DOI: 10.3390/s25010060.

[7] Novikov O., Ilin M., Stopochkina I., Duduladenko V. Stealthy Cyberattacks on Control Systems Using an Adaptive Soft-Constrained Optimization Method. Theoretical and Applied Cybersecurity. 2025. Vol. 7, No. 1. P. 104–110. DOI: 10.20535/tacs.2664-29132025.1.333440.

[8] ASTI Automation S.R.L. Smart Process Control Learning System (2 × Air Fan). User Manual. Code number AA900.010.04. Bucharest, Romania, 2024.
[9] Tkáčik T., Tkáčik M., Jadlovská S., Jadlovská A. Design of Aerodynamic Ball Levitation Laboratory Plant. Processes. 2021. Vol. 9, No. 11. Article 1950. DOI: 10.3390/pr9111950.
[10] Timmerman P., van der Weele J. P. On the Rise and Fall of a Ball with Linear or Quadratic Drag. American Journal of Physics. 1999. Vol. 67. P. 538–546. DOI: 10.1119/1.19320.
[11] Åström K. J., Hägglund T. Advanced PID Control. Research Triangle Park, NC: ISA, 2006. 460 p. ISBN 978-1-55617-942-6.
[12] Liu J., Zhang W., Ma T., Tang Z., Xie Y., Gui W., Niyoyita J. P. Toward Security Monitoring of Industrial Cyber-Physical Systems via Hierarchically Distributed Intrusion Detection. Expert Systems with Applications. 2020. Vol. 158. Article 113578. DOI: 10.1016/j.eswa.2020.113578.

Information security and monitoring of the functioning processes of local information and communication systems

Oksana Onyshchuk[1]

[1] *Lesya Ukrainka Volyn National University ,Voli 13, 43025 Lutsk, Ukraine*

Abstract

This article presents and explains the issues of information security and monitoring of the functioning processes of local information and communication systems. The capabilities of **Active Directory** for centralized management of users, computers, security policies, and access rights are analyzed. The functionality of the **Wazuh** platform for collecting, processing, and analyzing system logs, detecting anomalous activity, and responding to potential information security incidents is investigated. The main anomaly detection methods, including rule-based analysis, statistical analysis, machine learning, and hybrid approaches, are compared. The feasibility of a comprehensive approach based on the integration of Active Directory and Wazuh and the correlation of events from different sources is substantiated. It is determined that the use of a hybrid approach can improve threat detection accuracy and reduce the number of false positives.

Keywords

cybersecurity, Active Directory, Wazuh, monitoring, system logs, anomalies, threat detection, SIEM.

1. Introduction

Modern information and communication systems are characterized by a continuous increase in the volume of information, the number of connected devices, and the complexity of hardware and software infrastructure. At the same time, the development of digital technologies is accompanied by an increase in the number of cyber threats, including phishing attacks, malware, unauthorized access, account compromise, attacks on network services, and attempts to obtain elevated privileges. Under such conditions, traditional access control to information resources is insufficient. A comprehensive approach is required that combines centralized management of users and resources with continuous monitoring of information security events [1, 2].

This issue is particularly important in corporate networks, where a large number of workstations, servers, users, and network resources must operate as a single, manageable system. One of the key tools for centralized management of such infrastructure is the Active Directory directory service. It provides centralized storage of information about users, computers, groups, and other domain objects, as well as mechanisms for authentication, authorization, and security policy enforcement [3].

An important role in the operation of Active Directory is played by the **Domain Controller**, which verifies user and device credentials, manages access to resources, and ensures the application of Group Policies. Replication between domain controllers improves the availability and consistency of information, while integration with DNS enables the discovery of network

Corresponding author

onyshchuk.oksana@vnu.edu.ua (O. Onyshchuk)

0000-0002-8342-3011 (O. Onyshchuk)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

resources. At the same time, improper configuration of domain controllers, Group Policies, or access rights may create additional vulnerabilities [4, 5].

An important component of Active Directory security is the use of **ACL, DACL, and SACL** mechanisms. A DACL determines which users or groups are allowed or denied specific operations on objects, while a SACL provides auditing of access events. Thus, access control mechanisms can not only restrict user actions but also generate information required for further security analysis.

However, even a properly configured centralized management system cannot provide complete protection without continuous monitoring of security events. It is necessary to detect unusual activity, configuration changes, unauthorized access attempts, modifications of critical files, and other potentially dangerous actions in a timely manner. The **Wazuh** platform is an appropriate solution for this task [6].

Wazuh is designed to collect, process, and analyze data from endpoints, servers, and other components of information infrastructure. Wazuh agents collect logs and information about system activity, after which the data is transmitted to the manager for analysis. At the next stage, events can be indexed and displayed using appropriate visualization tools. This architecture makes it possible to centrally analyze a large number of security events [4-6].

Therefore, the main problem addressed in this study is the need to develop an effective security system that simultaneously provides centralized access management, configuration monitoring, user activity auditing, log collection, anomaly detection, and rapid incident response. Particular attention should also be paid to selecting the optimal method for analyzing system logs, since rule-based, statistical, machine-learning-based, and hybrid approaches have different advantages and limitations.

2. Scientific Novelty Compared with Existing Studies

In the scientific and technical literature, the functioning of **Active Directory** and **Wazuh** is primarily considered separately. Microsoft documentation and specialized sources describe authentication mechanisms, Group Policies, access control, Domain Controllers, and FSMO roles. In turn, Wazuh documentation and specialized publications focus on agent installation, rule configuration, log collection, event monitoring, threat detection, and incident response. The research presented in this work uses, in particular, Wazuh documentation, Microsoft Learn materials, resources on Active Directory administration, and R. Gupta's book *Security Monitoring with Wazuh*.

The scientific novelty of the proposed research does not consist in developing a fundamentally new security algorithm, but rather in providing a comprehensive consideration of the interaction between centralized management tools and security monitoring tools. Unlike approaches in which Active Directory is considered only as a directory service and Wazuh as an independent monitoring system, this study considers them as complementary components of a unified security system for corporate information infrastructure.

The proposed approach involves separating the functions between two levels. **Active Directory** is primarily responsible for the identification of users and devices, centralized access management, enforcement of Group Policies, and organization of domain objects. **Wazuh**, in turn, provides event collection and analysis, change monitoring, detection of suspicious activity, and support for incident response.

Another element of the scientific novelty is the systematization of anomaly detection methods for system logs. The study compares four approaches: predefined rule-based analysis, statistical analysis, machine learning, and a hybrid approach. It was established that rule-based analysis is characterized by simplicity and rapid response but has limited capabilities for detecting unknown

attacks. Statistical methods make it possible to identify deviations from normal behavior, but they may generate false positives. Machine learning provides greater adaptability; however, it requires a sufficient amount of data and more complex configuration. The hybrid approach combines the advantages of different methods and is considered a promising direction.

Thus, the proposed concept is based on a **multilevel approach** in which centralized access management is complemented by continuous monitoring and event correlation. This makes it possible to move from simple responses to individual recorded events toward the analysis of event sequences that may indicate complex cyberattacks (Table 1).

Table 1
Comparison of Anomaly Detection Methods in System Logs

Method	Operating Principle	Advantages	Disadvantages	Main Application
Rule-Based	Compares events against predefined rules and patterns	Easy to configure; quickly detects known threats	Poor at detecting new and modified attacks	Detection of known attacks and security policy violations
Statistical	Identifies deviations of current values from normal behavior	Can detect unusual activity without a specific attack pattern	May generate false positives when normal behavior changes	Analysis of workload, traffic, and user activity
Machine Learning	Analyzes large volumes of logs using classification and clustering algorithms	Highly adaptive; capable of detecting unknown threats	Requires training data and more complex configuration	Detection of complex and unknown anomalies
Hybrid	Combines rule-based, statistical analysis, and machine learning	High adaptability; improved accuracy; reduced false positives	More complex to implement; requires additional resources	Comprehensive monitoring of corporate infrastructure

3. Brief overview of the proposed solution

The proposed solution is based on the comprehensive integration of Active Directory and Wazuh within a local information and communication system.

The first component of the solution is Active Directory, which provides centralized management of the domain infrastructure, including users, groups, computers, Organizational Units and Group Policies. Organizational Units make it possible to structure domain objects according to departments, locations or other criteria and apply appropriate security policies. This approach simplifies centralized administration and supports the separation of administrative responsibilities.

The second component is access control. ACL and DACL mechanisms are used to define permitted and prohibited operations on system objects, while SACL provides auditing of relevant

activities. The combination of access control and auditing creates a foundation for continuous security monitoring. As a result, unauthorized or abnormal activities can not only be restricted but also recorded in system logs.

The third component is Wazuh. Agents installed on endpoints collect system logs and other information related to system operation and security. The collected data is transmitted to the Wazuh manager, where it is processed and analyzed using predefined and custom rules. The Wazuh architecture includes agents, a manager, indexing components and visualization tools, enabling centralized security monitoring.

An important feature of the proposed solution is the ability to create custom detection rules. Wazuh can be configured to monitor changes to critical files, detect suspicious connections and identify other events specific to the organization's infrastructure. This makes it possible to adapt the monitoring system to the particular security requirements of the organization.

To improve monitoring efficiency, centralized data storage and event visualization can be implemented. The proposed solution considers the integration of Wazuh with Elastic Stack and Kibana, which enables the creation of dashboards and facilitates analysis of the current security state of the system. Such integration improves the efficiency of administrators and simplifies the identification of suspicious events.

Another important component is proactive threat hunting. Instead of responding only to already generated alerts, proactive monitoring involves searching for indicators of potential attacks. Wazuh can analyze logs, attack indicators and information about adversary tactics, techniques and procedures. Integration with MITRE ATT&CK and Osquery can further expand the capabilities of threat analysis and system monitoring.

A multi-level approach to anomaly detection is proposed. The first level uses predefined rules to quickly identify known threats. The second level applies statistical methods to detect deviations from normal system and user behavior. The third level uses machine learning techniques to analyze large volumes of data and identify previously unknown behavioral patterns. The final level is a hybrid approach that combines the results of several detection methods.

A key feature of the proposed solution is event correlation across multiple data sources. For example, an individual failed login attempt may not represent a significant threat. However, a large number of failed login attempts combined with unusual resource usage, modifications to critical files or suspicious network connections may indicate a coordinated attack. Correlation of such events provides a more complete picture of the organization's security state.

Reducing false positives requires appropriate detection thresholds and regular adaptation of system configuration. This is particularly important for statistical and machine learning methods, whose results may be affected by changes in normal user or system behavior. A hybrid approach makes it possible to consider both known attack patterns and deviations from normal behavior, potentially improving detection accuracy and reducing the number of false alerts.

An additional practical consideration is system performance optimization. Excessive log collection, unnecessarily frequent monitoring and improper agent configuration may negatively affect endpoint performance. Therefore, the solution should prioritize relevant data sources, configure appropriate collection intervals, filter unnecessary information and implement suitable data-retention policies.

Proposed Local Information and Communication System (LICS): Active Directory (AD): manages users, groups, OU and GPO, providing access control through ACL/DACL/SACL; **Wazuh:** collects logs, monitors endpoints using agents and FIM, and analyzes security events; **Hybrid analysis:** combines rules, statistics and machine learning; **Event correlation:** identifies abnormal

activity and generates alerts; **Incident response:** alerts trigger appropriate actions to mitigate security incidents.

Thus, the proposed solution represents an integrated security system in which Active Directory provides centralized identity, access and policy management, while Wazuh provides continuous monitoring, event analysis, anomaly detection and support for incident response. The combination of these capabilities creates a foundation for a more resilient and effective information security system.

4. Conclusions

The conducted research provided a comprehensive analysis of information security and monitoring tools for local information and communication systems. The study demonstrated that Active Directory provides effective centralized management of users, computers, access rights and security policies. Wazuh complements these capabilities by enabling centralized collection, processing and analysis of system and security logs, continuous monitoring of endpoints, detection of suspicious activities and creation of custom security rules.

The obtained results confirm the feasibility of integrating Active Directory and Wazuh within a corporate network. Such integration combines centralized identity and access management with continuous security monitoring, allowing potentially malicious activities and unauthorized access attempts to be detected at an earlier stage. The comparison of detection methods showed that a hybrid approach combining rule-based, statistical and machine learning techniques can provide more effective anomaly detection and reduce the number of false positives.

Further development of the proposed solution should focus on improving event correlation, adaptive adjustment of detection thresholds, expanding the use of machine learning and automating incident response. The effectiveness of the system also depends on proper configuration, regular auditing and optimization of data collection. Overall, the integration of Active Directory and Wazuh provides a reliable technological foundation for building a flexible, scalable and effective information security syste

References

[1] Microsoft. Active Directory Domain Services Overview. Microsoft Learn. URL: https://learn.microsoft.com/en-us/windows-server/identity/ad-ds/

[2] Microsoft. Group Policy Overview. Microsoft Learn. URL: https://learn.microsoft.com/en-us/windows-server/identity/ad-ds/manage/group-policy/group-policy-overview

[3] Microsoft. Controlling Access to Objects in Active Directory Domain Services. Microsoft Learn. URL: https://learn.microsoft.com/en-us/windows/win32/ad/controlling-access-to-objects-in-active-directory-domain-services

[4] Wazuh. Wazuh Documentation: Architecture. Wazuh Documentation. URL: https://documentation.wazuh.com/current/getting-started/architecture.html

[5] Wazuh. Log Data Analysis. Wazuh Documentation. URL: https://documentation.wazuh.com/current/user-manual/capabilities/log-data-collection/log-data-analysis.html

[6] Wazuh. File Integrity Monitoring. Wazuh Documentation. URL: https://documentation.wazuh.com/current/user-manual/capabilities/file-integrity/

mazewall: restricting application workers on Linux

Leanid Piliptsevich[1]

[1] *Independent researcher*

Abstract

An application may need access to the network and filesystem even when one of its components needs neither. Giving every worker the same permissions exposes unnecessary resources when that component processes malicious input. This paper presents mazewall, a library that uses Linux security mechanisms to limit selected workers. A process-wide policy blocks operations that the whole application does not need. Additional worker policies restrict system calls and access to files. A document-processing service provides a running example of how these restrictions are chosen, installed, and checked. The current implementation uses Kotlin and targets the Java Virtual Machine, but the underlying controls belong to Linux and can support other language integrations. The paper explains the implementation, its relationship to earlier syscall-filtering systems, and the limits of protection inside a shared process. Existing demonstrations provide preliminary evidence; they do not establish complete protection against code execution or measured performance benefits.

Keywords

Linux security, syscall filtering, Seccomp-BPF, Landlock, least privilege

1. Statement of the problem in general terms

Consider a Linux service that receives an uploaded document, extracts its text, and sends the result to another service. Its request handler needs a network connection. Its document parser needs to read the uploaded file. Its delivery worker needs to send the result. The parser itself does not need to start programs, open new network connections, or read unrelated files. Nevertheless, those operations may be available to it because it runs inside the same application process.

This becomes dangerous when a document makes the parser do something unintended. For example, an XML document can ask a vulnerable parser to include the contents of a local file. A crafted input can also make vulnerable code contact an attacker-controlled server. Fixing the parser remains necessary. Additional operating-system restrictions can reduce the damage by denying the particular file access or connection, even when the parser attempts it.

A policy for the entire application cannot always express these differences. If the delivery worker legitimately needs networking, a blanket network ban would break the service. mazewall therefore combines restrictions for the whole process with stricter restrictions for selected workers. Here, a worker means an operating-system thread dedicated to a class of application tasks. A policy means a description of permitted or forbidden operating-system operations.

A process contains the application's memory and resources. Threads are execution paths inside that process. Two threads can run different tasks while still sharing data and open files. Giving one thread fewer permissions changes what it can request from Linux; it does not divide the process into separate machines. Keeping this distinction in view explains both the benefit and the limitation of mazewall: different workers can have different rules without paying the architectural cost of moving every component into a separate process.

Corresponding author

pilleo19@gmail.com (L. Piliptsevich)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The research question is practical: how can an application restrict a worker without breaking the runtime that executes it, and what protection does that restriction actually provide? The proposed model assumes that the Linux kernel, runtime, and installation code are trusted. It primarily addresses trusted code processing untrusted input. The current library runs on the Java Virtual Machine (JVM), but the problem and the Linux enforcement mechanisms are not specific to Java.

2. Scientific novelty compared to known works

mazewall builds on established Linux mechanisms. Elasticsearch already provides an example of process-level syscall filtering in a JVM service [1]. The contribution presented here is a library-level combination of process restrictions, dedicated-worker policies, filesystem controls, and installation rules. The document-processing example shows how the developer can assign different permissions within one application while accounting for the runtime and shared process state.

Earlier research also automates policy construction and specializes filters. Confine analyses container applications and dependencies to derive syscall policies [2]. SYSPART finds server workers' serving phases through binary analysis and inserts Seccomp installation code at the corresponding transitions [3]. These systems address how to discover and restrict operations required by an existing workload. mazewall gives the application developer an explicit point at which to assign a policy to selected workers.

SysComb tracks developer-specified execution states and uses eBPF to enforce state-dependent syscall filters without rewriting the application [4]. Its work confirms that different thread states can justify different syscall sets. mazewall uses persistent Linux restrictions on dedicated workers and adds Landlock for filesystem policy. It does not claim to introduce thread-specific filtering or to outperform these systems. Its proposed contribution is the practical composition of these controls behind application-facing interfaces, together with an explicit account of what each boundary protects.

3. Brief presentation of the proposed solution

A system call is a request from a program to the Linux kernel. Opening a file, creating a socket, and executing a program are examples. Linux checks these requests regardless of which source language produced them. mazewall uses two existing Linux facilities: Seccomp-BPF to restrict system calls, and Landlock to restrict access to resources such as files. These controls complement normal file permissions and any restrictions already imposed by the deployment environment.

Seccomp-BPF runs a small filter when a thread makes a system call. The filter can inspect the call number and numeric arguments, then allow the call or reject it. The kernel can return a permission error without performing the requested operation. A filter can distinguish operations such as creating a socket and executing a program. It cannot directly read a pathname stored at a pointer supplied by the application. Linux documentation consequently describes syscall filtering as one part of building a sandbox, rather than a complete sandbox by itself [5].

Landlock supplies the filesystem part of the policy [6]. For the document parser, a policy can permit reading a selected input directory while denying covered accesses elsewhere. The actual protection depends on the rights requested and those supported by the running kernel. Seccomp answers a question such as whether a worker may execute a program; Landlock can

answer whether it may open a particular file with a requested kind of access. The library needs both kinds of restriction for the example.

For instance, suppose documents are stored under /srv/inbox and an unrelated credential file is stored under /srv/secrets. The intended parser policy permits the required reads under /srv/inbox, together with files needed by the runtime, but excludes the credential file. An ordinary document can then be opened. If a malicious document makes the parser attempt to open the credential file, Linux checks the attempted access against the installed rules. The decision depends on the resource access, not on recognizing the malicious document's syntax or identifying the vulnerability that triggered it.

The first level is a process-wide baseline: restrictions that every part of the application must obey. If the service never needs to execute another program after startup, that behaviour can be blocked throughout the process. A narrower policy is then installed on the parser workers. It can prevent them from creating network connections and limit which files they can open, while the delivery workers retain the network access needed for their job.

The restrictions accumulate. An operation must survive the outer deployment controls, the process baseline, and the restrictions on the thread making the request. A worker cannot use its own policy to regain an operation already denied by the baseline. Likewise, an application cannot use mazewall to undo a restriction imposed by its container or host. This allows a deployment policy and an application policy to serve different purposes without granting extra permissions.

The running example is therefore concrete: the process must not execute another program; the parser may open allowed input files but must not create a network connection; the delivery worker may connect where its own policy permits. These are intended policy outcomes, not measurements from an experiment. They also concern operations made by the relevant worker. They do not imply that the parser has a separate memory space or that every possible way of communicating has disappeared.

3.1. Recording intended behaviour: SBoB and policies

The developer starts by identifying what the workload should do. For the parser, this includes the uploaded document, any required schemas, and runtime files needed while processing it. It excludes unrelated secrets, output directories, and network access. This step records intended behaviour. Observing what an application happens to do during one execution cannot decide which of those actions should be trusted.

This distinction is central to a Software Bill of Behavior (SBoB). A Software Bill of Materials lists what software contains; an SBoB describes what it is intended to do when it runs. The published SBoB specification, version 0.0.3, covers expected file access, child processes, network destinations, and Linux capabilities [7]. It is currently a proposed specification. In the running example, the intended behaviour includes reading uploaded documents and delivering results, but excludes reading credentials or executing another program.

SBoB provides a way to communicate that intent, while mazewall supplies Linux controls for enforcing compatible restrictions. For example, a declaration that no external program should be executed can guide a process-wide execution ban. Observed behaviour still needs developer review before becoming policy. This relationship does not imply that every SBoB field maps directly to Seccomp or Landlock, or that the current library implements the complete published format.

mazewall includes profiling facilities that help discover the operating-system operations used by a workload [8]. Representative executions can reveal required files and system calls that were not

obvious from the application code. The result is a candidate for review. Successful inputs alone are insufficient: malformed documents, missing files, recovery paths, and shutdown can exercise different operations. Dependency and runtime upgrades can change the required set as well.

After review, the developer validates the policy using both allowed and forbidden operations. A valid document should still be processed successfully. An attempted connection from the parser should fail under its network restriction. An attempted read outside the allowed filesystem area should fail under the applicable Landlock rules. A denial is useful only if the intended workload still works; a configuration that prevents every task from completing has not demonstrated useful containment.

Installation must finish before a worker receives untrusted input. In the current library, required Landlock setup happens before Seccomp installation because the final syscall filter can prohibit the setup operations themselves. A setup failure is reported rather than silently treated as a protected worker. Running without protection is a different deployment choice and must be explicit. The application also has to select restrictions compatible with the Linux features available in its environment.

Restrictions remain attached to the operating-system thread. Seccomp filters cannot be removed to make that thread unrestricted again. Consequently, a restricted parser worker should remain dedicated to compatible parsing tasks. It cannot simply return to a general pool where a later task expects broader permissions. Worker creation, successful installation, task admission, and termination are therefore part of the security design, alongside the policy rules themselves.

The application also needs a defined response to denial. If the parser receives a permission error while processing an uploaded document, it can reject that document and report the failure to its caller. Retrying the same operation on an unrestricted worker would defeat the intended policy. During development, the error may instead reveal a missing legitimate requirement. The developer should investigate and revise the policy deliberately, then rerun the allowed and forbidden cases. An unexpected failure is therefore evidence to examine, not an automatic reason to grant broader access.

3.2. Language support

mazewall currently uses Kotlin and the JDK Foreign Function and Memory API to call Linux interfaces. The same Linux controls can support bindings for other languages; those bindings are not yet demonstrated here.

Each integration must preserve operations needed by its runtime and install restrictions on the correct operating-system thread. Logical tasks such as coroutines may share or move between threads, so restrictions do not automatically follow them. The current JVM integration uses dedicated platform threads for installation.

3.3. What the restrictions protect, and what they leave open

The direct protection is straightforward: when a restricted worker requests a forbidden operation, the applicable Linux control denies it. Changing the language-level caller does not remove the installed filter. This can prevent a parser from opening a prohibited file or creating a prohibited connection. However, an application process contains more than the worker's future system calls. Shared memory, already-open resources, and other workers remain important.

First, threads share memory. A thread restriction does not create a private heap or prevent access to objects already available to the code. Second, a file descriptor is a handle to an opened file or socket. Preventing a worker from opening a new socket does not necessarily stop it from writing

through an existing one. A policy allowing ordinary writes for legitimate I/O may leave that route available. Resource ownership therefore has to be considered before interpreting a network or filesystem policy as complete isolation.

Third, execution can be delegated. If compromised code can submit a task to a less-restricted worker, the task runs with that worker's restrictions. The process baseline still blocks the operations covered by the baseline, but the original parser policy does not travel with the submitted task. For this reason, dedicated-worker containment is most useful for trusted code handling risky data. Arbitrary hostile code may require a separate process with carefully limited communication and resources.

The repository contains a historical report of eight vulnerable-application scenarios [8]. A protected request intended to make the server contact another endpoint reports a permission error. A malicious XML example no longer returns the tested local-file content. These observations are consistent with the intended restrictions, but several report entries mainly show application errors or missing callbacks. Such symptoms alone do not prove that the vulnerable operation was reached and then denied by the kernel.

The report also includes a database-query attack that succeeds with and without containment. This illustrates a useful limit: harmful application behaviour can occur entirely through operations that remain permitted. The records are preliminary observations from the existing JVM demonstration, not new experiments for this paper. They do not establish a general exploit-prevention rate, portability across runtimes, or performance overhead.

A stronger evaluation should run identical inputs without containment, with only the process baseline, and with the additional worker policy. Each run should preserve a successful benign control and independently check the intended effect: whether a connection reached its destination, a program executed, or a file changed. The policy, source revision, kernel, runtime version, architecture, and outer deployment restrictions should be recorded. Tests involving an existing descriptor and delegation to another worker would also check whether observed behaviour matches the stated limitations.

4. Conclusions containing main obtained results

mazewall applies Linux restrictions at two useful levels: a baseline for the application process and additional rules for dedicated workers. Seccomp limits system calls; Landlock adds filesystem access controls. The current Kotlin/JVM implementation demonstrates the integration approach, while the kernel model can guide bindings for other languages. Each binding must still preserve runtime progress and attach restrictions to the correct operating-system thread.

The main design result is a practical set of rules for using these controls together: define the workload's needs, review observations, install protection before accepting tasks, and keep restricted workers dedicated to compatible work. The analysis also explains why shared memory, existing handles, and delegation limit what protection inside one process can promise. Controlled experiments remain necessary to establish effectiveness and cost across workloads.

Declaration on Generative AI

OpenAI Codex assisted with source discovery, manuscript drafting, revision, and document preparation. The author is responsible for verifying the technical claims, references, and final text.

References

[1] Elastic, Bootstrap checks, System call filter check. https://www.elastic.co/guide/en/elasticsearch/reference/current/bootstrap-checks.html (accessed 7 September 2026).

[2] S. Ghavamnia, T. Palit, A. Benameur, M. Polychronakis, Confine: Automated System Call Policy Generation for Container Attack Surface Reduction, in: RAID 2020, USENIX Association, 2020, pp. 443-458. https://www.usenix.org/conference/raid2020/presentation/ghavanmnia.

[3] V. L. Rajagopalan, K. Kleftogiorgos, E. Göktaş, J. Xu, G. Portokalidis, SYSPART: Automated Temporal System Call Filtering for Binaries, arXiv:2309.05169v2, 2023. https://arxiv.org/abs/2309.05169.

[4] M. Rossi, M. Abbadini, M. Beretta, D. Facchinetti, S. Paraboschi, SysComb: Fine-Grained Transparent System Call Filtering for Attack Surface Reduction, preprint, arXiv:2608.26871v1, 2026. https://arxiv.org/abs/2608.26871.

[5] The Linux kernel contributors, Seccomp BPF (SECure COMPuting with filters). https://docs.kernel.org/userspace-api/seccomp_filter.html (accessed 7 September 2026).

[6] The Linux kernel contributors, Landlock: unprivileged access control. https://docs.kernel.org/userspace-api/landlock.html (accessed 7 September 2026).

[7] Bill of Behavior, SBOB specification, version 0.0.3, proposed specification, 2026. https://billofbehavior.com/ bob/docs/spec/ (accessed 8 September 2026).

[8] L. Piliptsevich, mazewall source repository, implementation guidance, and vulnerable-application demonstration report. https://github.com/Pilleo/mazewall (local checkout inspected 7 September 2026; reported experiments not independently reproduced).

Digital-Twin-Referenced Physical-Layer Anomaly Detection for a 1-Bit RIS in 6G Networks

Andriy Riy[1], Marian Kyryk[1], Serhii Zablotskyi[1] and Ivan Demydov[1]

[1] *Lviv Polytechnic National University, S. Bandery St. 12, 79013 Lviv, Ukraine*

Abstract

Reconfigurable intelligent surfaces (RIS) make terahertz 6G links viable but move part of the control plane to the physical layer, where conventional intrusion detection does not reach. We reuse the radio digital twin, already required for beam management, as a security reference, evaluated by ray tracing at 280 GHz against the geometry and 1-bit control of a nanoporous-alumina RIS under design. The twin proves decisive: the same network reaches 0.809 AUROC over five attacks with it and 0.670 without. The relational model is not: across five seeds a 1234-parameter MLP autoencoder and a 24516-parameter ST-GNN are indistinguishable, 0.812 ± 0.010 against 0.809 ± 0.008, the autoencoder sixty times faster. We also report the governing constraint — at 280 GHz a twin accurate to centimetres cannot predict phase — and an active 1-bit probe that survives it.

Keywords

6G, terahertz communication, RIS digital twin, physical-layer security, anomaly detection

1. Problem Statement

The terahertz band is a promising candidate for ultra-wideband 6G access, but it imposes severe propagation and hardware constraints: path loss grows with the square of frequency, and quasi-optical propagation makes links acutely sensitive to blockage [1]. One remedy is to control the environment with RIS panels that restore non-line-of-sight connectivity [2, 3], with digital twins predicting channel state from geometry [4].

That creates a security problem. Once beam selection and RIS configuration follow reported channel state information (CSI), the radio environment becomes an attack surface: an adversary need not break cryptography, only manipulate the evidence the control loop consumes, and higher-layer detection reacting in seconds is hopeless when the channel decorrelates in 453 μs. Five attacks define the problem: a CSI report replaced by the eavesdropper's channel at unchanged power (A1); pilot contamination (A2); a hijacked phase matrix that focuses the panel elsewhere while network and twin believe the commanded profile (A3); a rogue panel co-located with the legitimate one, so only phase structure distinguishes it (A4); and a fabricated blockage report (A5).

All five create a mismatch between the measured physical behaviour of a link and what the twin predicts for it, yet none is separable by a fixed per-link threshold, because the residual between reality and any twin is itself non-zero and mobility-dependent.

2. Scientific Novelty

Physical-layer security for RIS has addressed secrecy, not configuration integrity [7]; the closest standards anchor [6] assumes a cellular control plane, not a programmable reflector; and digital twins in 6G are treated as a performance asset [4].

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

- The twin is used not to accelerate beam management but as a verifiable physical reference for security, so no new sensing infrastructure is needed — the reference already exists.
- Phase is excluded from the feature vector by necessity rather than by choice: at 280 GHz a twin accurate to centimetres is wrong by tens of wavelengths. All residuals are therefore phase-insensitive, and the passive comparison is complemented by active 1-bit probing of the panel.
- We evaluate against a panel actually being built, so its 1-bit aperture loss and mirror lobe sit in the channel rather than assumed away.

3. Proposed Solution

Ray tracing gives the predicted channel as the direct path plus the cascade: the RIS-to-terminal channel weighted by the phase profile, times the base-station-to-RIS channel.

3.1. What a terahertz twin cannot predict

Configuring the panel from the twin and comparing measurement against it does not survive the physics. At 280 GHz a twin localising terminals to 5 cm is wrong by some forty-seven wavelengths, so its per-element phase estimates are effectively random and any coherent residual would measure localisation error rather than an attack.

The architecture must therefore be inverted: the network aligns the panel by beam training on the measured channel, and the twin predicts the outcome using its own optimal alignment for the same user. Both being aligned, they agree in every phase-insensitive observable up to twin geometry error, and the residual vector is built from those: path gain, dominant-beam distance, first-arrival delay, a phase-invariant subspace mismatch and the challenge residual.

3.2. Challenge–response on a 1-bit panel

The network perturbs a cryptographically seeded quarter of the panel and the twin predicts the resulting power change. It survives the phase blindness because an aligned panel's amplitude is the sum of element-wise magnitudes, so the response depends only on the illumination taper. A replayed report is static and shows no response, so forgery needs a secret the attacker lacks.

On a 1-bit panel the perturbation is a bit flip, because the pump can only toggle a cell. The measured response is 3.50 dB on the served link against 1.57 dB for a continuous dither, and 0.00 dB in line of sight, so the probe authenticates RIS-served links specifically.

3.3. Detectors

Residuals form a dynamic base-station–RIS–user graph scored by a masked spatio-temporal graph autoencoder with neighbourhood, temporal and bottleneck heads. All detectors train on benign traffic only, so every attack is unseen by construction. Heteroscedastic scoring is required: the benign residual scale varies 21-fold across links, so an unnormalised error follows the noisiest.

4. Evaluation and Results

The proposed detector is the masked ST-GNN on twin residuals. Five reference detectors each remove one ingredient: an MLP autoencoder (no graph, no memory), a recurrent model (no graph), a threshold on the twin residual (no learning), a CUSUM test on path gain alone (no twin features), and the same ST-GNN denied the twin. The comparison isolates what each contributes.

The scenario is generated with Sionna RT [5] on the simple_street_canyon scene at 280 GHz: a 64-element base station, the panel on a building corner, and twelve pedestrians at 0.4–2 m/s, half in line of sight and half in a cross street served through the RIS (Figure 1). Snapshots are 20 ms apart, 300 per realisation, three realisations — two for training, one held out.

The attacker may forge reports, transmit pilots, overwrite the applied phase profile and deploy a co-located reflector, but cannot alter what the base station measures or learn the probe seed.

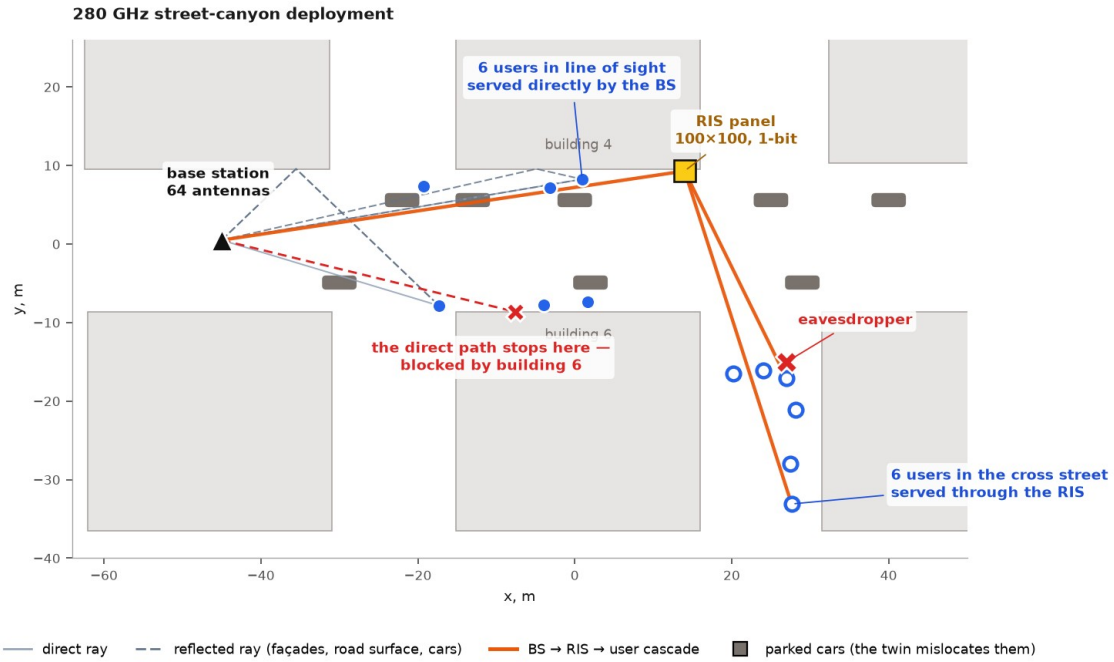


Figure 1: The evaluated deployment and its traced propagation paths.

The panel is a device, not a parameter. Its geometry and control are taken from a photoexcited-silicon nanoporous-alumina RIS under design: a 35 × 35 mm aperture on a 350 µm lattice, hence 100 × 100 = 10 000 optically switched elements. These are design values — no S-parameters or scattering patterns of a fabricated panel exist yet, so everything reported here is simulation-based evidence.

Control is 1-bit; its two costs — $(2/\pi)^2$ in aperture efficiency and a mirror beam at $-\theta$ — emerge from the quantisation rather than being applied as penalties, at −3.97 dB in simulation against −3.92 dB predicted. Beyond about 1.0 dB of double-pass window loss the panel is idealised as a lossless two-state reflector with exact addressing. Finite phase precision, frequency dependence, mutual coupling, addressing errors, thermal drift and the microsecond latency of optical switching are outside the model, and each would degrade the figures below.

Dielectric surfaces use the ITU-R P.2040 [8] glass parameters for 220–450 GHz, the only ones defined that high.

The twin has the exact static geometry but a degraded dynamic scene: cars displaced by about 0.6 m, one missing, terminals located to ±5 cm, leaving a benign path-gain residual of −0.07 ± 3.91 dB — unbiased but wide.

The median line-of-sight link sits at −92.8 dB against −133.3 dB for the served link's cascade, and the cascade-served users lack a direct path in 56 % of snapshots, so the panel is their only link more than half the time.

Attacks are injected as episodes of 4–12 snapshots covering 40 % of the held-out trace, 20 realisations each, labelled per link.

Displacement rather than deletion is deliberate: dropping the cars removes the reflections that carry energy around the corner, giving a twin that misses the serving link for half the users.

Figure 2 maps the cascade power on a 0.25 m grid. Hijacking the phase matrix moves the focus 18 m onto the eavesdropper and strips 20.6 dB from the served link, while the commanded configuration — and hence the twin's prediction — is unchanged.

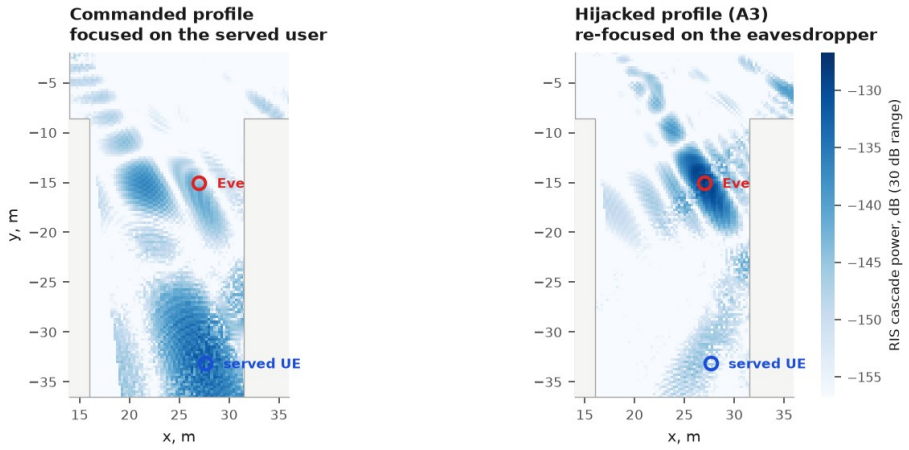


Figure 2: Cascade power through the panel for the commanded phase profile (left) and a hijacked one (right).

Table 1
Detection on the held-out trace, macro-averaged over the five attacks

Detector	AUROC	FPR	Episodes	Params	Latency, ms
CUSUM on path-gain residual	0.597	0.000	0.00	2	<0.01
ST-GNN, no digital twin	0.670 ± 0.008	0.086	0.47	24 516	0.61
Digital twin, no GNN	0.710	0.016	0.30	10	<0.01
LSTM, no graph	0.766 ± 0.004	0.030	0.35	12 010	0.03
Masked ST-GNN + digital twin	0.809 ± 0.008	0.039	0.50	24 516	0.62
MLP autoencoder + digital twin	0.812 ± 0.010	0.028	0.47	1 234	<0.01

The twin supplies most of the detection power. The identical ST-GNN denied the twin reaches 0.670 against 0.809, and per attack it barely beats chance on the RIS hijack, 0.51 against 0.82. Capacity is not the lever either: the single-statistic CUSUM test consumes the same residual and reaches 0.597, flagging no episode at all. All rates are per link and per snapshot: FPR is the fraction of benign link-snapshots above the calibrated threshold, Episodes the fraction of injected episodes flagged before they end, and latency the time to score one snapshot of all twelve links.

The relational model does not pay for itself. Across five training seeds the MLP autoencoder and the ST-GNN are indistinguishable, 0.812 ± 0.010 against 0.809 ± 0.008; over the 5675 injected episodes the paired difference is +0.003, 95 % CI [−0.011, +0.017], p = 0.57. Repeating the comparison on an idealised continuous-phase panel — the ablation that would expose any advantage the 1-bit hardware removes — leaves them so too, 0.822 ± 0.003 against 0.822 ± 0.012. The ST-GNN never held an advantage for the hardware to remove, and it costs twenty times the parameters and sixty times the latency. Per attack they do differ: the autoencoder leads where the evidence sits in one link, the ST-GNN where a sustained deviation betrays the attack. A single run would have suggested otherwise, which is why every learned result is reported over five seeds.

Twin fidelity decides which detector is appropriate. With exact positions a plain threshold on the twin residual is best in the study, 0.856 at 0.9 % false positives. It collapses to 0.708 at the realistic 5 cm operating point, and at 20 cm nothing works.

5. Limitations and Future Work

The panel is idealised: its design analysis has the phase contrast falling from 180° to 115° for a 2 % detuning at $Q \approx 15$, so a fabricated device will do worse. Realised false-positive rates exceed the 1 % target for every learned detector, which is why AUROC is used for comparison and why lowering them is the first priority. Measured S-parameters from a fabricated panel, real CSI traces, wider mobility, multi-user attacks and the imperfections listed above remain open.

6. Conclusions

A digital twin built for beam management can serve a second purpose: the reference against which physical reality is continuously verified. The governing constraint is that at terahertz frequencies it cannot predict phase, so residuals must be phase-insensitive — a restriction that still admits an active probe by flipping pixels. Against a panel actually being built, the twin reference carries the detection accuracy, while the relational model is indistinguishable from an MLP autoencoder sixty times faster. These are proof-of-concept results from simulation, not evidence of deployment readiness: false-positive rates of 2.8–8.6 % remain well above the 1 % target, and lowering them is the principal open problem.

Acknowledgements

This research received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101086493 — HORIZON-MSCA-2021-SE-01 (project “TeraHertz”) and from the LASER-PRO project.

Declaration on Generative AI

During the preparation of this work, the authors used a generative AI assistant in order to: drafting assistance, code implementation for the simulation campaign, and grammar and spelling checking. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] I. F. Akyildiz, C. Han, S. Nie, Combating the distance problem in the millimeter wave and terahertz frequency bands, IEEE Communications Magazine 56 (2018) 102–108.
[2] C. Pan et al., Reconfigurable intelligent surfaces for 6G systems: Principles, applications, and research directions, IEEE Communications Magazine 59 (2021) 14–20.
[3] K. Dovelos, S. D. Assimonis, H. Q. Ngo, B. Bellalta, M. Matthaiou, Intelligent reflecting surfaces at terahertz bands: Channel modeling and analysis, in: Proc. IEEE ICC Workshops, 2021.
[4] L. U. Khan, W. Saad, D. Niyato, Z. Han, C. S. Hong, Digital-twin-enabled 6G: Vision, architectural trends, and future directions, IEEE Communications Magazine 60 (2022) 74–80.
[5] J. Hoydis et al., Sionna RT: Differentiable ray tracing for radio propagation modeling, arXiv:2303.11103, 2023.
[6] 3GPP, Study on 5G security enhancements against false base stations (FBS), TR 33.809, 2023.
[7] J. Chen, Y.-C. Liang, Y. Pei, H. Guo, Intelligent reflecting surface: A programmable wireless environment for physical layer security, IEEE Access 7 (2019) 82599–82612.
[8] ITU-R, Effects of building materials and structures on radiowave propagation above about 100 MHz, Recommendation ITU-R P.2040-3, 2023.

Modeling Critical States In Cyber System Elements Based On $In_xGa_{1-x}Sb_yAs_{1-y}$ Alloy

Hennadii Shapovalov[1] and Anatoly Kazakov[1]

[1] *Odessa National Polytechnic University, Shevchenko Avenue, 1, Odessa, Ukraine*

Abstract

Modeling of processes of phase coexistence spaces formation in $In_xGa_{1-x}Sb_yAs_{1-y}$ solid solution was performed. Thermodynamic functions for description of states of binary compounds and interaction of atoms in quaternary solid solutions were used. Higher derivatives of free energy of homogeneous solid solution $In_xGa_{1-x}Sb_yAs_{1-y}$ with mixing in two sublattices from the third to the eighth derivative inclusive were calculated. Calculations of derivatives within the framework of regular solution model and standard thermodynamic functions were performed. Sections of phase diagrams were calculated using analytical expressions for higher derivatives. Positions of zero contours of higher derivatives were determined based on differential topological approach. Sections of phase diagram of $In_xGa_{1-x}Sb_yAs_{1-y}$ solid solution, critical spaces and phase coexistence spaces for different temperatures in four-dimensional concentration space were calculated. Calculation results for the possibility of forming second-order phase coexistence spaces in $In_xGa_{1-x}Sb_yAs_{1-y}$ solid solutions indicate that the resulting simulations are consistent with available experimental data. The possibility of forming phase coexistence spaces in materials used in modern cyber systems may lead to malfunctions under critical operating conditions.

Keywords

phase coexistence spaces, differential topological approach, solid solutions, multicomponent systems, cyber system elements

1. Introduction

Four-component solid solutions based on compounds of groups III–V are currently promising materials to produce cyber system elements. Such elements, obtained based on epitaxial heterojunctions, are intended for operation in a wide spectral range. The four-component solid solution $In_xGa_{1-x}Sb_yAs_{1-y}$ was chosen as the object of study in the work. Its physical properties under critical operating conditions are currently little studied. However, this material has attracted the attention of researchers as a promising material for the creation of optoelectronic devices with a wide spectral range, efficient LEDs, and high-speed photodiodes that can be used to solve cybersecurity problems. Its application in the field of security is associated with ensuring the physical protection of objects and the protection of information systems. Materials based on the In-Ga-Sb-As system make it possible to create detectors operating in high-resolution mode (SWIR/MWIR) at room temperature [1]. Such detectors are used in security systems for perimeter monitoring, night vision and object recognition [2]. These sensors are used in early threat detection systems, gas leak detection, and hidden thermal anomaly detection. This use of sensors prevents sabotage or man-made accidents [3]. The use of infrared light makes it possible to create biometric identification devices (vein scanners, IR-facial recognition). Such devices are more difficult to deceive than standard video cameras, ensuring access security [4]. The investigated materials are used for hardware in the field of cybersecurity and information protection. Their application provides physical layer protection and threat detection using extended spectral analysis [5]. IR-sensors based on In-Ga-Sb-As allow the creation of cameras operating in the short-wave infrared

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

(SWIR) range. They can distinguish living human skin from a silicone mask, photo or display. This is critical for access control systems and protection against spoofing. SWIR cameras on In-Ga-Sb-As material provide a clear image of the subcutaneous vein pattern. This method is a more secure biometric marker than a fingerprint. The material studied in the work is used in secure video surveillance systems (Machine Vision) for monitoring in conditions of limited visibility. Such systems provide control over critical infrastructure at night, in fog or in the presence of smoke screens, detecting potential intruders. Detectors built on In-Ga-Sb-As are used in devices for physical protection of communication channels. Their application is associated with the tasks of detecting interference in monitoring optical networks for unauthorized connection or attempts to remove information by detecting leakage of IR radiation [6, 7]. In general, the use of the In-Ga-Sb-As system strengthens the technical basis of cybersecurity and increases the reliability of the physical level of data and infrastructure protection.

However, the solid solution $In_xGa_{1-x}Sb_yAs_{1-y}$ loses stability under certain conditions [8 – 14] because of spontaneous ordering. In this regard, research into the self-organized formation of ordered structures in the In-Ga-Sb-As system with a non-uniform distribution of component concentrations is of considerable interest. Predicting the properties of materials under critical operating conditions is the goal of such research. Using computer modeling to analyze the processes leading to the emergence of self-organized, ordered structures, multidimensional phase diagrams can be calculated, considering the possible existence of bifurcation spaces, critical spaces, and spaces where phases of different orders coexist.

A thermodynamic approach to describing phase transitions in the In-Ga-Sb-As system, which allows for the analysis of the properties of potential functions of several order parameters, enables the use of catastrophe theory for self-organizing systems. Thom's catastrophe theory, as a generalized form of the Ginzburg-Landau phase transition theory, is used in this work [15]. An equation of state on some n-dimensional manifold in the corresponding space is constructed. The state of the system in which one stable state coexists with another stable state is the criterion for defining phase coexistence spaces. The process of emergence of such a space is a first-order phase transition according to Maxwell's principle. The potential function will then have two or more global minima of equal depth. Bifurcation subspaces, in which a stable phase can become unstable, can arise within the phase space under study under certain conditions. Second-order critical spaces can arise in regions where two different phases become identical. Critical spaces of order 3 and 4 are formed in the presence of three or four identical phases, respectively.

2. Purpose of the study

The aim of the work is to predict critical states in the quaternary solid solution $In_xGa_{1-x}Sb_yAs_{1-y}$ based on the differential topological approach [15] and thermodynamic modeling. Higher derivatives of potential functions of several order parameters from the third to the eighth were obtained for calculating the phase coexistence spaces. Calculations within the framework of the regular solution model considering the mixing of atoms in two sublattices were performed [16].

3. Critical Space Model

An analysis of the quaternary solid solution $In_xGa_{1-x}Sb_yAs_{1-y}$ as a system in which the existence of multiple phases was assumed was carried out based on a generalization of the conditions for the existence of stable phases [15] to the case of a concentration space with a

dimension of four. This approach made it possible to verify the possibility of the emergence of critical spaces and phase coexistence spaces in the solid solution under consideration.

The Gibbs free energy of the studied solid solution in the considered model as a mixture of four hypothetical binary compounds was considered:

$$dG = G(X_{AC},\ X_{BC},\ X_{AD},\ X_{BD}), \tag{1}$$

where X_{ij}– concentrations of binary components. Expressions for the concentrations of binary components under the assumption of a random distribution of heterogeneous atoms over the nodes of the corresponding sublattices take the following form [16]:

$$X_{AC} = (1-x)(1-y), \qquad X_{BC} = x(1-), \qquad X_{AD}(1-x)y, \qquad X_{BD} = xy, \tag{2}$$

where x and y – concentration parameters.

Analytical expressions for higher derivatives of the potential energy of the system with respect to the order parameters were obtained and investigated to test the possibility of occurrence of critical spaces and phase coexistence spaces in the solid solution under consideration. Higher derivatives from the third to the eighth for the Gibbs free energy (1) of the solid solution $In_xGa_{1-x}Sb_yAs_{1-y}$ with respect to the concentration parameters were calculated by matrix-vector differentiation of multidimensional systems. The direct sum method for obtaining matrices of higher derivatives was used [17, 18].

The calculation of the higher derivatives of the free energy of the system under consideration was carried out according to the following algorithm. The expressions for the second derivatives μ_ij of the Gibbs free energy of the solid solution $In_xGa_{1-x}Sb_yAs_{1-y}$, which are stable to differentiation, obtained in the work [19] within the framework of the regular solution model were used. The matrix G_2 of dimension 4×4 from the expressions of the second derivatives of the free energy (1) μ_{ij} was formed:

$$G_2 = \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} \\ \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} \\ \mu_{31} & \mu_{32} & \mu_{33} & \mu_{34} \\ \mu_{41} & \mu_{42} & \mu_{43} & \mu_{44} \end{pmatrix} \tag{3}$$

Next, matrix G_2 was differentiated by the four corresponding concentrations of X_{ij}. Four 4x4 matrices of third-order partial derivatives were obtained because of differentiation. The block-diagonal matrix of the total third derivative of the free energy of the system under study was formed in the next step using the direct sum method:

$$G_3 = \begin{pmatrix} \frac{dG_2}{dX_{AC}} & 0 & 0 & 0 \\ 0 & \frac{dG_2}{dX_{AD}} & 0 & 0 \\ 0 & 0 & \frac{dG_2}{dX_{BC}} & 0 \\ 0 & 0 & 0 & \frac{dG_2}{dX_{BD}} \end{pmatrix} \tag{4}$$

The determinant obtained matrix G_3 as the product of the determinants of the matrices of partial derivatives located on the main diagonals was determined:

$$\det G_3 = \det G_{13} \times \det G_{23} \times \det G_{33} \times \det G_{43}, \quad (5)$$

where

$$G_{13} = \frac{d G_2}{d X_{AC}}; G_{23} = \frac{d G_2}{d X_{AD}}; \quad (6)$$
$$G_{33} = \frac{d G_2}{d X_{BD}}; G_{43} = \frac{d G_2}{d X_{BC}} \ .$$

Obtaining derivatives of the free energy of the system of higher orders from the fourth to the eighth inclusive and the corresponding determinants $\det G_i$, where i is the order of the derivative (i=1, 2, 3,... , 8), was carried out using a similar algorithm. Analytical expressions of higher derivatives for calculating the corresponding determinants of matrices were obtained.

The positions of the zero contours of the derivatives under consideration were calculated after obtaining analytical expressions for the determinants of the matrices of derivatives of the free energy of the system under consideration. The values of the concentration parameters x and y in the range from 0 to 1 with a step of 0.1 (x, y $\in (0;1)$) were substituted into the expressions for the corresponding determinants for this purpose. The fulfillment of the condition $\det G_i = 0$ was checked at the next step. The found points on the cross section of the existence of solid solutions in the phase diagram were displayed. Next, the areas in which the values of the corresponding determinants take positive and negative values were determined. Positive values were marked with light areas, and negative values with darker areas.

4. Simulation results

The locations of the regions of existence of derivatives and their zero contours from the second to the eighth derivative inclusive for the temperature range of 773K–1023K were determined. The zero contours of the analytical expressions obtained in this work for the total derivatives of the Gibbs free energy of the quaternary solid solution $In_xGa_{1-x}Sb_yAs_{1-y}$ on the corresponding section of the state diagram of the system under study were numerically determined and plotted. The thermodynamic parameters used in the calculations are presented in Table 1.

Table 1.
Parameters of interaction in the solid phase for quasi-binary systems

Quasi-binary system	Interaction parameter α^5_{ij-jk}, J/mol, [20]
GaAs-InAs	12570
GaSb-InSb	7961
GaAs-GaSb	18855
InAs-InSb	9427,5

The results of calculating the positions of the zero contours of the derivatives of the Gibbs free energy in the $In_xGa_{1-x}Sb_yAs_{1-y}$ system based on concentration parameters, starting with the second derivative and up to and including the eighth derivative, for a temperature of 773 K are presented in Figures 1 – 7.

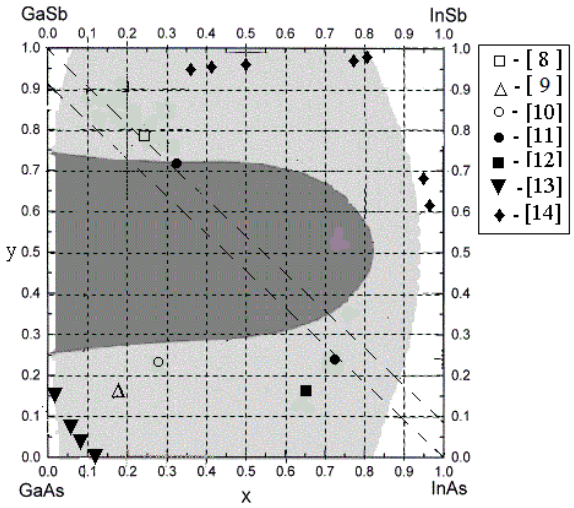


Figure 1: The results of numerical calculations of the second derivative of free energy in the $In_xGa_{1-x}Sb_yAs_{1-y}$ system on the cross-section of the existence of solid solutions of the state diagram are presented.

The space of negative derivative values is shown in dark colors, and the space of positive derivative values is shown in lighter colors. The isoperiodic lines for the *GaSb* and *InAs* substrates are shown as dashed lines in Figures 1 – 4. Experimental compositions of $n_xGa_{1-x}Sb_yAs_{1-y}$ epitaxial layers grown on *GaSb* and *InAs* substrates are shown in the diagrams. Single-crystal layers on a *GaSb(111)B* substrate were grown by liquid-phase epitaxy in [8].

Satisfactory quality of the layers for the case of positive values of the lattice mismatch parameter was noted. Modulation of the composition of layers grown on a *GaSb (100)* substrate with a period of about *150nm* was observed in [9]. Modulation acquired a more clearly expressed periodicity with increasing values of x and y. The composition obtained in [10] was grown on a GaSb (111) substrate.

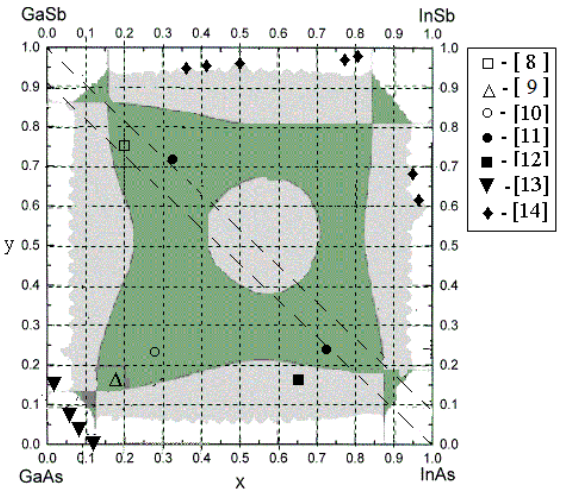


Figure 2: The results of numerical calculations of the third derivative of free energy in the $In_xGa_{1-x}Sb_yAs_{1-y}$ system on the cross-section of the existence of solid solutions of the state diagram are presented.

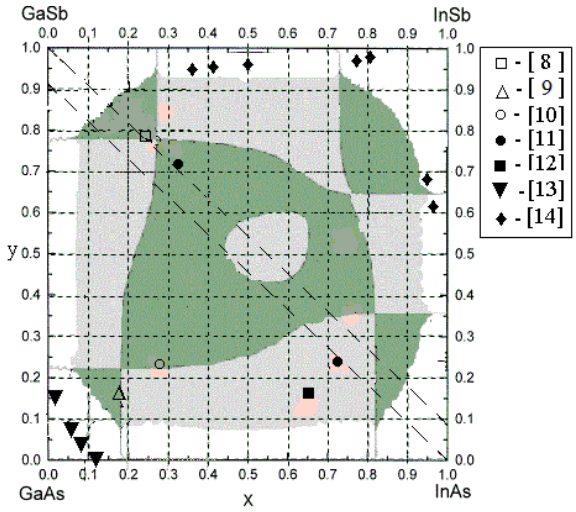


Figure 3: The results of numerical calculations of the fourth derivative of free energy in the $In_xGa_{1-x}Sb_yAs_{1-y}$ system on the cross-section of the existence of solid solutions of the state diagram are presented.

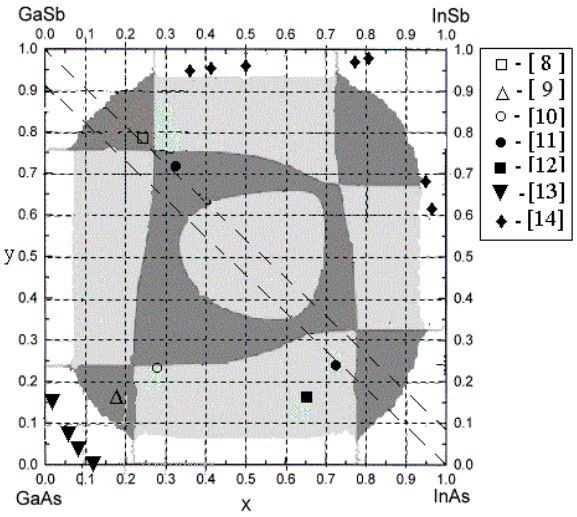


Figure 4: The results of numerical calculations of the fifth derivative of free energy in the $In_xGa_{1-x}Sb_yAs_{1-y}$ system on the cross-section of the existence of solid solutions of the state diagram are presented.

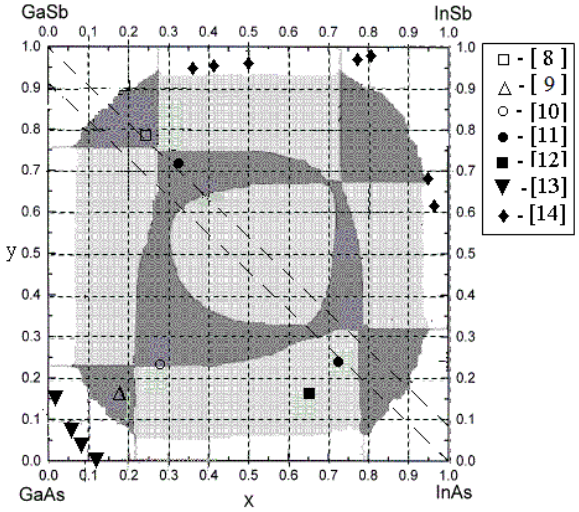


Figure 5: The results of numerical calculations of the sixth derivative of free energy in the $In_xGa_{1-x}Sb_yAs_{1-y}$ system on the cross-section of the existence of solid solutions of the state diagram are presented.

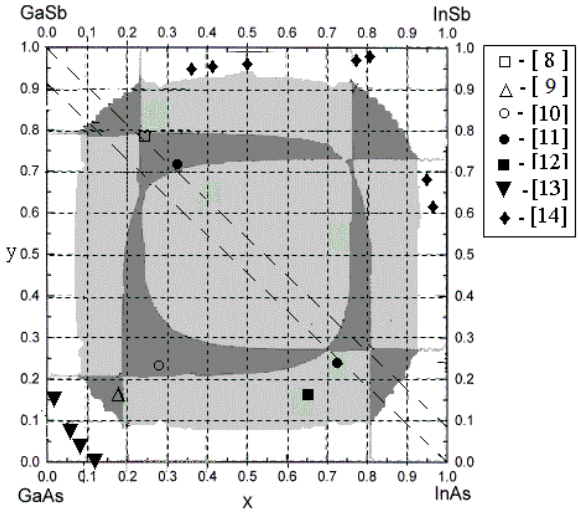


Figure 6: The results of numerical calculations of the seventh derivative of free energy in the $In_xGa_{1-x}Sb_yAs_{1-y}$ system on the cross-section of the existence of solid solutions of the state diagram are presented.

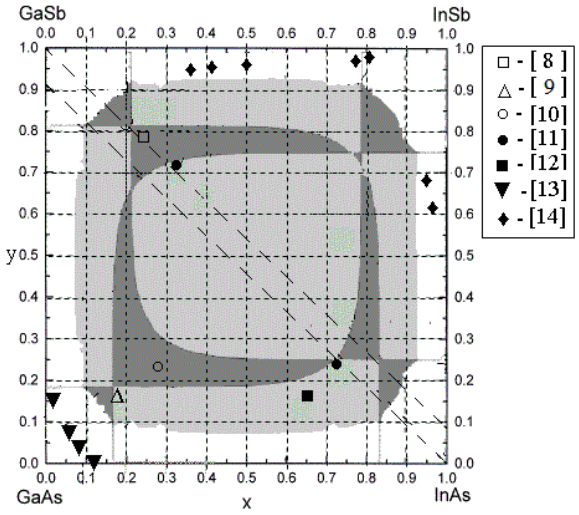


Figure 7: The results of numerical calculations of the eighth derivative of free energy in the $In_xGa_{1-x}Sb_yAs_{1-y}$ system on the cross-section of the existence of solid solutions of the state diagram are presented.

The compositions corresponding to the experimental points [11] were obtained near the immiscibility region and are not stable. The epitaxial layers described in [12] were obtained using LPE and MOVPE technologies under conditions close to thermodynamic equilibrium. The points at which the growth of single-crystal [13] and polycrystalline layers [14] were observed on *GaSb* and InAs substrates, the initial growth compositions for which were in the region of thermodynamic instability, were also marked in the figures.

The contours of the phase coexistence regions of order two were constructed based on the obtained contours of the third and fourth derivatives on the cross-section of the solid solution phase diagram of the system under study. The results of modeling the phase coexistence spaces on the cross-section of the solid solution phase diagram of the $In_xGa_{1-x}Sb_yAs_{1-y}$ system at a temperature of 773K are presented in Figure 8.

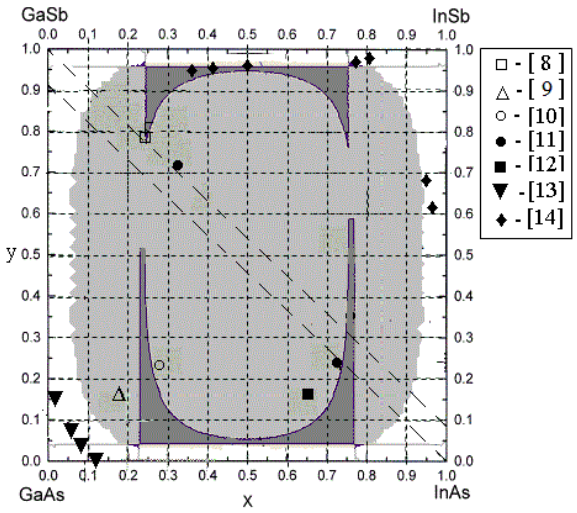


Figure 8: The results of numerical calculations of the coexistence spaces of phases of order two in the $In_xGa_{1-x}Sb_yAs_{1-y}$system are presented.

The spaces of fulfillment of the conditions for the formation of phase coexistence spaces of order two, where the coexistence of two solid phases is possible, are shown in dark colors. The space of existence of the third and fourth derivatives is shown in lighter colors. The results of numerical calculations of the derivatives of the free energy of the system from the second to the fourth inclusive, presented in Figs. 1 – 4, were used to construct these regions. The data on the experimental composition [8], as can be seen from Fig. 8, fall into the phase coexistence space. The point corresponding to the composition [9] is in the immediate vicinity of the calculated space. Most experimental compositions [10-12] are in the immediate vicinity of the calculated space of coexistence of two phases. Compositions [13, 14], due to their clearly expressed single-crystal and polycrystalline structure, do not fall into the space of existence of derivatives.

5. Conclusions

The results of modeling the coexistence spaces of order two phases in the $In_xGa_{1-x}Sb_yAs_{1-y}$ system obtained in this work allow us to explain the available experimental data on the spontaneous formation of two-phase modulated structures in the epitaxial layers of the considered solid solution. It is shown that the used regular solution model for the Gibbs free energy allows for a sufficiently correct assessment of the position of the coexistence space of order two on the section of phase diagrams. The obtained modeling results allow us to assume that the proposed model for predicting the space of possible occurrence of three-dimensional modulated structures in layers of the $In_xGa_{1-x}Sb_yAs_{1-y}$ solid solution can also be used. The model considered in this work for cases of occurrence of three coexisting phases for different compositions of the considered solid solution by constructing critical spaces of order three can also be used.

References

[1] S.A. Nelson, J.M. Fastenau, D. Lubyshev, M. Kattner, Frey P., Liu A.W.K., Martinez B., Furlong M.J. Large Format Multi-Wafer Production of LWIR Photodetector Structures on 150mm GaSb Substrates by MBE / SPIE Defense, Commercial Sensing Conference, Anaheim, CA, Apr 26-30, 2020, 11407-38.

[2] D. Lubyshev, J.M. Fastenau, M. Kattner, P. Frey, S.A. Nelson, R. Flick, M. Rogers, A.W.K. Liu, P. Flint, and N. Faleev. Effect of Substrate Orientation on Sb-based MWIR Photodetector Characteristics / Infrared Phys. Technol. 95, 27 (2018).

[3] A. Liu, D. Lubyshev, J.M. Fastenau, S. Nelson, M. Kattner, and P. Frey. Molecular Beam Epitaxial Growth and Characterization of Large-format GaSb-based IR Photodetector Structures / OSA Opt. Mater. Express, 8, 1282 (2018).

[4] Masatoshi Koyama, Junichi Kudo, Yasuhiro Iguchi, Hiroshi Inada, Yasuhiro Muramatsu, Isamu Kumihashi, Masahiro Sano. Indium Gallium Arsenide/Gallium Arsenide Antimonide Type II Superlattice Photodetector That Can Utilize Night Airglow / Sensors and Materials, Vol. 36, No. 6 (2024) 2183–2197, MYU, Tokyo.

[5] A. Mostefai. Basic Characteristics of Gallium Indium Arsenide Antimonide ($Ga_xIn_{1-x}As_ySb_{1-y}$) Semiconductors / Journal Of Nano-And Electronic, Vol. 14, No 4, 04027(4pp) (2022).

[6] Yang Zhang, Lidan Lu, Yang Chen, Leidong Shi, Weiqiang Chen, Lianqing Zhu. Multispectral infrared detectors based on photonic crystal plates / Optics Express, Vol. 33, No. 2 / 27 Jan 2025 / 3373.

[7] F. Mohammedy, M. Jamal Deen. Growth and fabrication issues of GaSb-based detectors / J. Mater Sci: Mater Electron (2009) 20:1039–1058.

[8] A. Joullie, F. Jia Hua, F. Karouta, H. Mani. LPE growth of GaInAsSb/GaSb system: The importance of the sign of the lattice mismatch / J. Crystal Growth 75 (1986) 309.
[9] C. A. Wang. Correlation between surface step structure and phase separation in epitaxial GaInAsSb // Appl. Phys. Lett. – 2000 – Vol. 76 – P 2077.
[10] V. Vasil'ev, A. Deryagin, V. Kuchinskii, V. Smirnov, G. Sokolovskii, D. Tret'yakov, N. Faleev. Properties of GaInAsSb solid solutions obtained from antimony fluxes by liquid-phase epitaxy in the spinodal decay region / Tech. Phys. Lett. 24 (1998) 231.
[11] M.J. Cheng, G.B. Stringfellow, D.W. Kisker, A.K. Srivastova, J.L. Zyskind. GaInAsSb metastable alloys grown by organometallic vapor phase epitaxy / Appl. Phys. Lett. 48 (1986) 419.
[12] A.N. Baranov, Y. Cuminal, N. Bertru, C. Alibert, A. Joullie. Strained multiple-quantum-well lasers grown on GaSb emitting between 2 and 2.4 um / Proc. SPIE 2997 (1997), 2.
[13] A.I. Kazakov, I.N. Kishmar, A.I. Bochkarev, L.V. Dolginov. Stability analysis of quaternary $In_xGa_{1-x}Sb_yAs_{1-y}$ alloys // J. Crystal Growth – 1992. – Vol. 116. – P.P. 204-212.
[14] Dolginov L. M., Eliseev P.G., Lapshin A.N., Milvidskii M.G. A study of phaseequilibria and heterojunction in Ga-In-As-Sb quaternary system / Kristall Tech. – 1978. – Vol. 13. – P. 631
[15] Okada, K. Classical calculations on the phase transition I. Phase diagram in four-dimensional space for the system with one order parameter / K. Okada, I. Suzuki // J. Phys. Soc. Jap. – 1982. – Vol. 51, # 10. – PP. 3250-3257
[16] Onabe, K. Thermodynamics of the type $A_{1-x}B_xC_{1-y}D_y$, III-V quaternary solid solutions. / K.Onabe // J. Phys. Chem. Solids, 1982 –V.43, #11, – PP.1071-1086
[17] A.I. Kazakov, G.V. Shapovalov, P.P. Moskvin. Computer simulation for formation of critical spaces in II–VI solid solutions / Journal of Crystal Growth, 2019, V. 506, pp. 201 — 205.
[18] Traat, I. Matrix calculus for multivariate distributions / I.Traat // Transactions of Tartu State Univ., – 1986. – Vol. 733. – PP. 64 – 85.
[19] Kazakov, A. Computer simulation for stability of quaternary solid solutions / A. Kazakov, I. Kishmar // J. Crystal Growth. – 1991. – Vol. 110. – PP. 803-814
[20] Streingfellow, G.B. / J. Phys. Chem. Solids, 1973, v. 34, N. 11, p. 1749 – 1751

GP4-WLTE 4G Controller Backdoor as an Example of Unauthorized Access Risks in Chinese Equipment

Mykola Shcherbyna[1] and Petro Venherskyi[1]

[1] *Ivan Franko National University of Lviv, Universytetska 1, 79000 Lviv, Ukraine*

Abstract

This work examines hidden backdoors in Chinese-manufactured devices, using the popular GP4-WLTE-EA/EC controller as a case study. Using hardware and software reverse engineering, an undocumented backdoor was discovered that discloses the access password via SMS, enabling full control over the device. Transitioning to a self-hosted MQTT broker instead of the vendor's cloud is proposed as a more secure alternative. Finally, the study demonstrates why such Chinese-manufactured IoT devices must not be deployed in critical infrastructure.

Keywords

4G, backdoor, Chinese, cloud, CWE-200, CWE-259, CWE-288, CWE-798, CWE-912, cyberattack, cybersecurity, GP4-WLTE, GSM, Internet of Things, IoT, MQTT, Remote Terminal Unit, reverse engineering, RTU, SMS, unauthorized access, undocumented command, war dialing

1. Introduction

The rapid proliferation of Internet of Things (IoT) devices has fundamentally transformed modern infrastructure. However, the balance between economic efficiency and system security has reached a critical inflection point. Unvetted commercial hardware deployed at scale creates an expanded attack surface.

The low cost of Chinese hardware products has driven their widespread adoption globally, yet this affordability is frequently accompanied by recurring cybersecurity incidents. A concerning pattern of critical vulnerabilities – such as hardcoded credentials and undocumented management interfaces – spans diverse product categories. Notable examples include hardcoded credentials in JAIOTlink IP cameras (CVE-2026-58453 [1]), vulnerabilities in TP-Link Wi-Fi routers (CVE-2026-12001 [2], CVE-2024-57040 [3]), and remote exploitation vectors in Diqee robotic vacuum cleaners (CVE-2018-10987 [4]). Together with numerous similar cases, these incidents highlight a dangerous trend in COTS (commercial off-the-shelf) device security.

While Chinese domestic cybersecurity legislation includes formal provisions aligned with European and American data protection standards [5], it exists alongside national security laws that implicitly permit or compel the inclusion of backdoors for state intelligence purposes [6]. The operational risks of such hardware opacity were further underscored by the recent controversy in Scandinavian countries concerning Yutong electric buses, which were found to incorporate remote telemetry modules capable of executing an off-site engine shutdown [7].

Amid the ongoing Russo-Ukrainian War, the security of cyber-physical systems and supply chains in critical infrastructure becomes a paramount national security concern [8]. In this high-threat environment, deploying unverified equipment in energy, municipal, or industrial networks introduces severe vulnerabilities that hostile actors can exploit [9].

Corresponding authors

mykola.shcherbyna@lnu.edu.ua (M. Shcherbyna); petro.venherskyy@lnu.edu.ua (P. Venherskyi)

0009-0003-9761-9466 (M. Shcherbyna); 0000-0001-9808-7404 (P. Venherskyi)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Motivated by these considerations, we decided to evaluate the security of popular Chinese IoT hardware. A widely sold GP4-WLTE controller (Figure 1) was bought on AliExpress e-commerce platform for hardware and firmware investigation [10]. As detailed in the following sections, our initial security concerns proved to be fully justified.

2. Unauthenticated Backdoor Discovery

2.1. Target Device Overview and Specifications

The GP4-WLTE-EA/EC 4G controller (Figure 1) is functionally classified as a wireless Remote Terminal Unit (RTU). Conceptually, it serves as an equivalent to industrial terminals such as the ICP DAS GTP-541M/GTP-542M series [11], while offering a cost that is an order of magnitude lower. The device is powered by a direct current (DC) supply within the 9…24 V range.

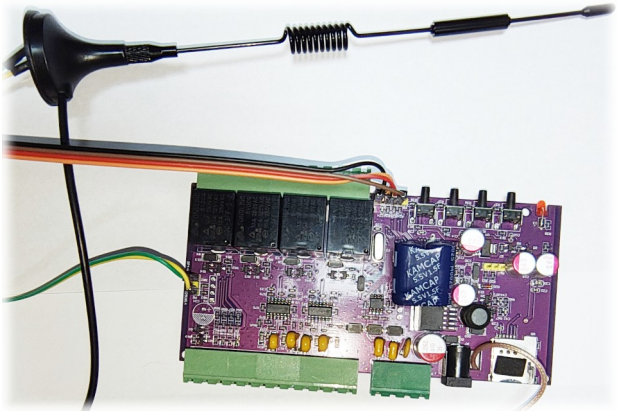


Figure 1: Exposed PCB of the GP4-WLTE 4G controller under investigation. [CC BY 4.0], via the authors.

The device features three hardware interface types:

- Analog Inputs (AI): 4 independent channels for receiving 4…20 mA current signals with a 12-bit analog-to-digital conversion resolution.
- Digital Inputs (DI): 4 channels supporting state detection of passive switches (“dry contact” type).
- Digital Outputs (DO): 4 relay channels with a switching capacity of 10 A at 250 V AC (or 10…15 A at 30 V DC). Each relay output is equipped with a dedicated LED indicator for visual monitoring.

The primary configuration tool is the “WLTE Control” mobile application [12], which allows users to set operational parameters and define complex automation logic, linking input readings directly to the relay outputs (DO) at the hardware level.

As an alternative communication channel, basic digital output control is implemented via Short Message Service (SMS). This channel is also utilized for dispatching SMS alarms – for example, when an analog signal (such as a fluid level reading) crosses a specified threshold. Initial controller provisioning is similarly conducted via SMS, primarily to configure APN parameters. For the Vodafone Ukraine mobile operator, the configuration command follows the format:

```
SN1234APN08=internet,00=0,00=0
```

where the APN name, username, and password are specified sequentially (comma-separated), with the two digits preceding the "=" sign indicating the character length of each respective field [13].

2.2. Hardware and Firmware Reverse Engineering

The device under investigation was disassembled for analysis. Reverse engineering revealed that it is built on a Chinese Geehy APM32FEPSCBT6 microcontroller (Figure 2). Documentation for this specific model is limited, indicating it is likely a variant intended for the domestic Chinese market, while the Geehy APM32 product family generally comprises ARM Cortex-M microcontrollers. Based on physical indicators (pin count, pinout, and naming suffix), the authors hypothesized that the core hardware is functionally identical to the APM32F103CBT6 – an ARM Cortex-M3 microcontroller compatible with the widely used STM32F103 series from STMicroelectronics [14].



Figure 2: Close-up view of the device's Geehy APM32FEPSCBT6 MCU in an LQFP-48 package. [CC BY 4.0], via the authors.

Firmware extraction from embedded systems is typically complicated by active Flash Readout Protection (RDP), sometimes requiring advanced hardware techniques such as Voltage Glitching, Clock Glitching, or Electromagnetic Fault Injection (EMFI) [15]. In this instance, however, readout protection was omitted, allowing the firmware image to be dumped using standard debug interfaces via an ST-Link V2 in-circuit debugger/programmer.

SHA-256 (firmware dump):

`1408F5544A7884D504E31DE0C79F6EABDC8F7CB21CD543C0F9137924CDEB549B`

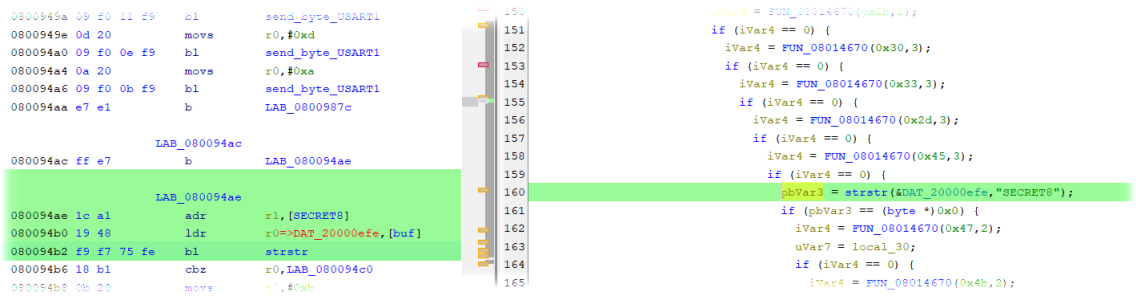


Figure 3: Comparison of backdoor command string in the Ghidra decompiler. [CC BY 4.0], via the authors.

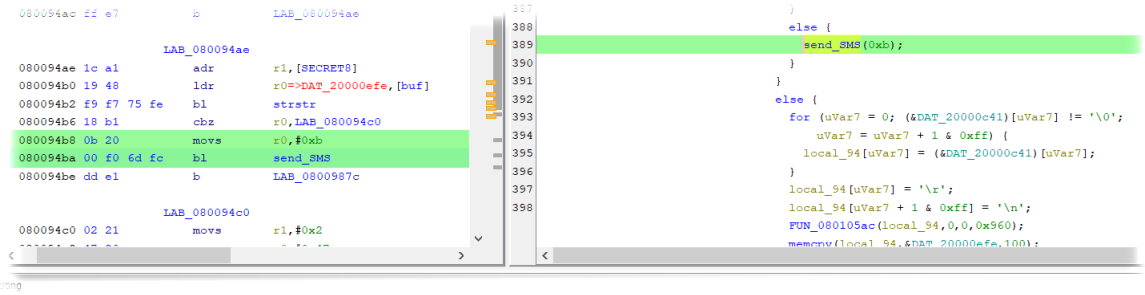


Figure 4: Function call passing a specific argument for the backdoor command, viewed in the Ghidra decompiler. [CC BY 4.0], via the authors.

Decompilation of the extracted binary was carried out in Ghidra [16]. By tracing documented control commands, an incoming SMS handler function was identified at address `0x08008C94`. The function routine contained an undocumented condition checking for the string "`SECRET8`" (Figures 3 & 4).

Upon matching this command, a response construction function at `0x08009D98` is invoked with parameter `0xB` (Figure 5), which subsequently calls the dispatch handler at `0x080103F4`. This subroutine generates an automated reply payload containing the active device password (Figure 6).

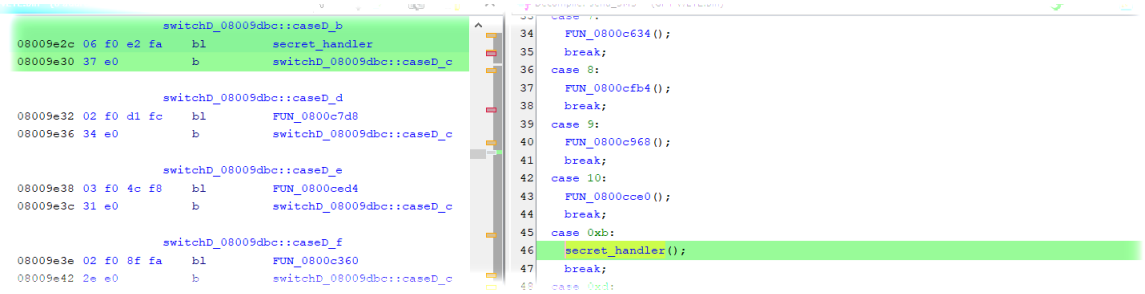


Figure 5: A *switch* statement calling the backdoor command handler in the Ghidra decompiler. [CC BY 4.0], via the authors.

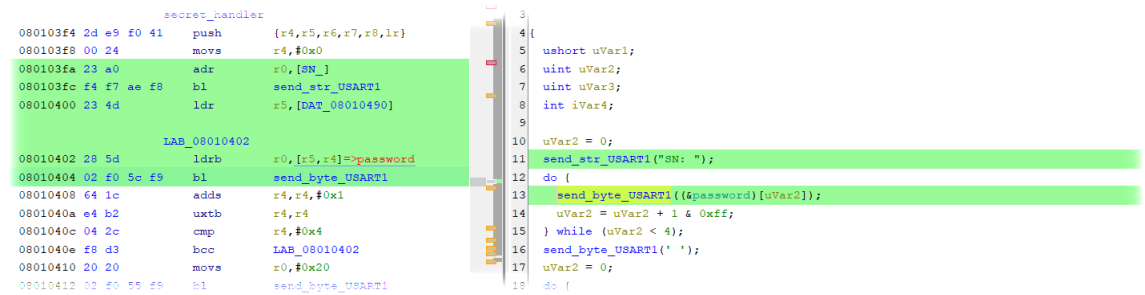


Figure 6: The backdoor command handler sending the current device password, viewed in the Ghidra decompiler. [CC BY 4.0], via the authors.

2.3. Experimental Backdoor Verification

To practically validate the vulnerability, test payloads containing the "`SECRET8`" command were transmitted with both valid ("`0000`") and invalid ("`7777`") password fields. In both cases, the device disclosed the active system password. Furthermore, the password was successfully reset to "`1234`", granting full remote control over the relay outputs (Figure 7).

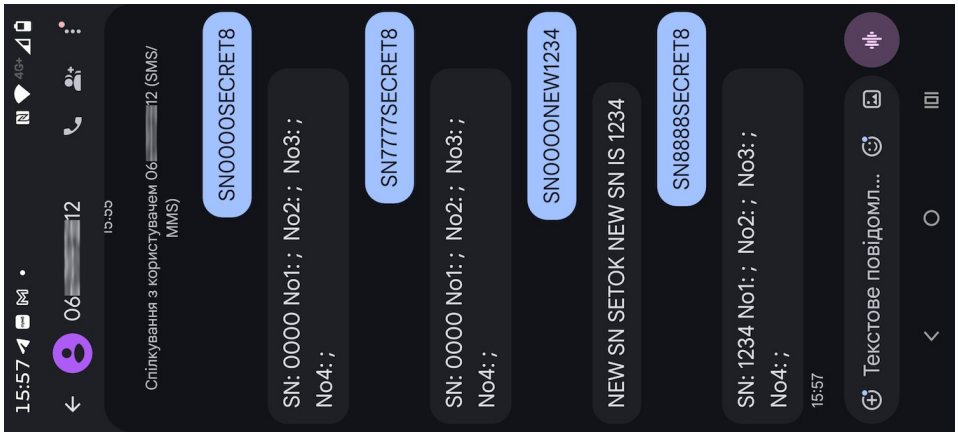


Figure 7: Test SMS session demonstrating password retrieval via a backdoor and its modification by a hacker. The phone number is blurred. [CC BY 4.0], via the authors.

Although the firmware allows registering "master" telephone numbers, these entries are enforced solely for voice-call "commands" authorization and are bypassed during SMS control processing. Consequently, we revealed a persistent, unauthenticated backdoor that can be exploited by an adversary aware of the target device's phone number.

2.4. CWE Mapping

The conducted analysis revealed the following common firmware weaknesses in the GP4-WLTE device [17]:

- CWE-200: Exposure of Sensitive Information to an Unauthorized Actor.
- CWE-259: Use of Hard-coded Password.
- CWE-288: Authentication Bypass Using an Alternate Path.
- CWE-798: Use of Hard-coded Credentials.
- CWE-912: Hidden Functionality.

This highlights the critical need for secure coding practices in IoT device development.

3. MQTT Protocol Functionality

Although the documentation for our GP4-WLTE sample made no mention of MQTT protocol [18] support, the ASCII string "MQTT" appeared repeatedly within the binary firmware dump, and several third-party vendor web pages referenced such an operational mode. Through further reverse engineering and experimentation, we deduced the format for the corresponding SMS configuration command:

SN*password*$_1$**MQTT***broker*,*port*,*timeout*,*user*,*password*$_2$

where *password*$_1$ is the current device access code, *broker*:*port* specifies the MQTT server IP address or domain alongside the target port, and *user*:*password*$_2$ represents the authentication credentials for the MQTT broker.

Of note, the firmware contains hardcoded references to a default server (`mqtt.svnwi.com`) and credentials (`user`:`1234`), which likely serve as a test MQTT broker. Since the *broker* field in the

MQTT configuration SMS command has a 32-character maximum length restriction – thereby making impossible the use of longer public cloud broker domain names – we deployed a dedicated own MQTT broker and provisioned its parameters into the controller.

Upon sending the MQTT configuration SMS, the device replied with a confirmation message enumerating the active connection parameters along with the status notification: "`Please wait while the network restarts`". Following a restart, the device successfully initiated an outbound connection and published telemetry to our custom broker. For instance, manually toggling relay state using physical button SW2 triggered the following JSON payload transmission:

```
{
 "deviceId": "reXXXXXXXX77",
 "msgType": "MANUAL_CONTROL",
 "date": "2026-XX-XX XX:XX:XX",
 "data": {
  "relays": [
   { "id": 1, "state": "OFF" },
   { "id": 2, "state": "ON" },
   { "id": 3, "state": "OFF" },
   { "id": 4, "state": "OFF" }
  ],
  "dryContactInputs": [
   { "id": 1, "state": 0 },
   …
  ],
  "analogInputSignal": [
   { "id": 1, "state": "offline" },
   …
  ]
 }
}
```

(the timestamp and the `deviceId` have been sanitized for privacy).

Notably, the target phone number is omitted from outbound MQTT payloads, reducing phone-number disclosure risks associated with the previously described SMS backdoor. Furthermore, standard cloud connectivity and MQTT operating modes appear to be mutually exclusive: enabling MQTT disables mobile app management (via the WLTE Control application) and web interface management (svnwi.com), restricting remote control exclusively to the custom MQTT telemetry channel, but still supporting the SMS command protocol.

4. Conclusion

In this study, we investigated the popular, low-cost GP4-WLTE-EA/EC wireless Remote Terminal Unit and discovered an undocumented backdoor activated via an SMS command. This vulnerability returns the current access password, enabling an attacker to fully compromise and intercept control of the device. Exploiting this flaw requires only knowledge of the target controller's mobile phone number, which an adversary must acquire through out-of-band methods. However, the classic "war dialing" attack vector remains highly viable. If a facility deploys a large fleet of these devices using bulk-purchased SIM cards with sequential phone numbers, the compromise of a single node could facilitate a cascading attack against the entire deployment.

The firmware's failure to enforce SMS filtering based on registered “master” numbers constitutes a critical security design flaw. Mitigating this attack vector is currently only possible at the cellular network operator level, requiring the implementation of strict SMS blocking policies for the specific SIM cards in use. For a comprehensive security assessment, future work should involve intercepting and analyzing the proprietary cloud communication protocol – or conducting further reverse engineering – to determine whether the device leaks its phone number to the cloud infrastructure. Regardless, the vendor's cloud platform inherently possesses unrestricted administrative control over connected devices.

For a more secure deployment, installers should utilize the MQTT mode, which empirically appears to decouple the controller from the vendor's cloud ecosystem. However, verifying the absolute isolation of this mode requires further in-depth firmware reverse engineering.

Ultimately, the GP4-WLTE-EA/EC serves as yet another example of low-cost hardware harboring undocumented remote access backdoors. In the geopolitical context of the Russo-Ukrainian War, such issues can no longer be viewed as mere vendor oversights. Rather, they function as latent cyberweapons capable of enabling targeted sabotage against critical energy, municipal, and industrial infrastructure. Despite their economic appeal, this controller and similar opaque IoT devices must be strictly prohibited from deployment within critical infrastructure environments. Only rigorously verified commercial solutions or auditable open-source architectures can provide the necessary security assurances for industrial IoT applications.

Declaration on Generative AI

During the preparation of this work, the authors used Google Gemini in order to translate text and improve writing style. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] CVE-2026-58453: JAIOTlink C492A-W6 4.8.30.57701411 Hard-coded Credentials via anyka_ipc, 2026. URL: https://www.cve.org/CVERecord?id=CVE-2026-58453.

[2] CVE-2026-12001: Hardcoded Credential Vulnerability in Multiple TP-Link Router Models, 2026. URL: https://www.cve.org/CVERecord?id=CVE-2026-12001.

[3] CVE-2024-57040: Hardcoded Password Vulnerability in TP-Link TL-WR845N Routers, 2024. URL: https://www.cve.org/CVERecord?id=CVE-2024-57040.

[4] CVE-2018-10987: Authenticated Remote Code Execution Vulnerability in Dongguan Diqee Diqee360 Vacuum Cleaners, 2018. URL: https://www.cve.org/CVERecord?id=CVE-2018-10987.

[5] Standing Committee of the National People’s Congress, Data Security Law of China, 2021. URL: https://www.chinajusticeobserver.com/law/x/data-security-law-of-the-people-s-republic-of-china20210610.

[6] M. S. Tanner, Beijing’s new National Intelligence Law: From defense to offense, Lawfare, 2017. URL: https://www.lawfaremedia.org/article/beijings-new-national-intelligence-law-defense-offense.

[7] Euronews, Chinese-made buses in Norway can be halted remotely, spurring increased security, 2025. URL: https://www.euronews.com/next/2025/11/06/chinese-made-buses-can-be-halted-remotely-in-norway-spurring-increased-security.

[8] Y. Khlaponin, Security threats when integrating cyber-physical systems into global networks, in: Proceedings of the International Scientific and Technical Conference “Security of Modern

Information and Communication Systems" (SMICS-2025), Lviv, Ukraine, 2025, pp. 129–132. URL: https://smics2025.lnu.edu.ua/wp-content/uploads/2025/11/2025-smics_zbirnik_final.pdf.

[9] O. Polevod, M. Shelest, Y. Tkach, Cybersecurity from within: how embedded backdoors undermine trust in national and international systems, in: Proceedings of the International Scientific and Technical Conference "Security of Modern Information and Communication Systems" (SMICS-2025), Lviv, Ukraine, 2025, pp. 34–39. URL: https://smics2025.lnu.edu.ua/wp-content/uploads/2025/11/2025-smics_zbirnik_final.pdf.

[10] AliExpress, GP4-WLTE Smart 4G Global GSM SMS APP Controller RS485 WEB Control, 2026. URL: https://www.aliexpress.com/item/1005006284883131.html.

[11] ICP DAS, Wireless Communication 3G/4G/5G Products: Remote Terminal Unit, 2026. URL: https://www.icpdas.com/en/product/guide+Wireless__Communication+3G_4G_5G__Products+Remote__Terminal__Unit.

[12] Google Play, WLTE control, 2026. URL: https://play.google.com/store/apps/details?id=com.ruiensi.rf.hongykzjv3activitys.

[13] 4G/LTE Controller. GP4-WLTE Intelligent Controller, 2026. URL: https://drive.google.com/file/d/1qNRIvq6eaBG2n22ksYR5YEi64AD5CNre/view?usp=drive_link.

[14] Geehy Semiconductor Co., Ltd., APM32F103 Series Microcontrollers, 2026. URL: https://global.geehy.com/product/fifth/APM32F103.

[15] J. van Woudenberg, C. O'Flynn, The Hardware Hacking Handbook: Breaking Embedded Security with Hardware Attacks, No Starch Press, 2021.

[16] National Security Agency, Ghidra: Software Reverse Engineering Framework, 2026. URL: https://github.com/NationalSecurityAgency/ghidra.

[17] MITRE, Common Weakness Enumeration (CWE), 2026. URL: https://cwe.mitre.org/index.html.

[18] MQTT: The Standard for IoT Messaging, 2026. URL: https://mqtt.org/.

Integration of CGAN and Federated Learning to Protect Decentralized SDN Networks from Adaptive EW

Maksym Stoyko[1] and Mykola Gnidenko[1]

[1] *Department of Computer Science, DUICT*

1. Introduction

In the author's previous research, the effectiveness of a decentralized subject-oriented SDN architecture was developed and substantiated, where Edge AI and Federated Learning (FL) provide the formation of "collective immunity" of tactical networks to interference. However, the experience of modern combat operations and the analysis of research in 2026 indicate the evolution of electronic warfare (EW) means. Static broadband jamming has been replaced by Adaptive Multi-Strategy Jamming (AMSJ) [2].

2. Statement of the problem

Modern AMSJ systems operate as autonomous agents (often based on the mathematical apparatus of a multi-armed bandit), dynamically combining strategies: bandwidth fragmentation, frequency spreading, unpredictable power bursts and narrowband smart jamming [2]. This generates non-stationary interference patterns that are difficult for AI models to distinguish from legitimate signals (I/Q samples) [1]. Classical Federated Learning algorithms (e.g. FedProx) face the problem of extreme local data imbalance, since nodes of a tactical network cannot exchange raw spectrograms for further training due to strict radio masking requirements (OPSEC) [3].

3. Research objective

Improving the mathematical model of distributed SDN-control of heterogeneous networks (SDN-ACS) by integrating generative adversarial networks (CGAN) into the Federated Learning process for fast classification and adaptive interference counteraction in non-IID distribution conditions.

4. Presentation of the main material

To solve the problem of data imbalance on isolated Edge nodes (UAV modules, command and control machines), an extended Fed-AugCGAN (Federated Augmented Conditional GAN) framework is proposed, which is deployed in the local Knowledge Plane of a decentralized SDN controller.

1. Local Data Generation (Edge AI): Instead of transmitting raw telemetry from the air to the headquarters, each Edge node locally applies the Windowed Fourier Transform (STFT) to extract spectrograms from the air. When encountering a new type of AMSJ interference, the local generative network (CGAN) synthesizes valid-like samples of this interference, artificially expanding the data set for training the local classifier without involving the real air [1, 2].

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Validation-Guided Aggregation: During gradient exchange, the SDN cluster coordinator performs Validation-Guided Weighted Aggregation. The weight of each device's contribution is dynamically adjusted based on the reliability of its local model. This allows the system to operate robustly in heterogeneous non-IID environments, eliminating "poisoned" data from compromised nodes.
3. Self-Healing SDN Routing: After accurately classifying the interference waveform (Waveform Classification), the local SDN controller automatically applies network countermeasures [3]. Using the principles of Intent-Based Networking, the network independently redistributes priority traffic (e.g., UAV swarm telemetry) to protected mesh channels or satellite links, leaving noisy spectrum for background tasks.

5. Conclusions

The integration of generative adversarial models (CGAN) into the Federated Learning loop allows Edge nodes of decentralized SDN networks to autonomously recognize and block the most advanced adaptive electronic warfare (AMSJ) jammers. The proposed hybrid architecture completely eliminates the problem of AI model degradation due to the lack of training samples at the front-line positions, while avoiding the unmasking exchange of raw radio signals. The implementation of Edge AI based on Fed-AugCGAN turns tactical SDN networks into self-organized "living" ecosystems capable of maintaining connectivity even during intense multi-domain countermeasures.

References

[1] *Meftah, A., Do, TN, et al.* (2026). Federated Learning-Enabled Jamming Detection and Waveform Classification for Distributed Tactical Wireless Networks. *IEEE Transactions on Network and Service Management.*

[2] *Wang, Y., et al.* (2026). CGAN-Augmented Federated Learning for Adaptive Multi-Strategy Jamming Detection in Stochastic Tactical Air–Ground Network. *IEEE Communications Letters, 30*, 1-5.

[3] AI-Driven Tactical Communication and Networking in Defense: The Future of Connected Battlespaces. (2026). *Electronics Era / Defense Technologies, 2.*

Algorithms for Accelerating Modular Operations in Cryptographic Systems

Iryna Vinkovska[1] and Lidiia Tymoshenko[1]

[1] *Odesa Polytechnic National University, 1 Shevchenka Ave., 65044 Odesa, Ukraine*

Abstract

The approach to optimizing modular multiplication and modular exponentiation using the Rademacher–Chrestenson number-theoretic bases and the residue number system (RNS) has been investigated. Multiplication using precomputed coefficients for pairs of digit positions and selective accumulation of active coefficients is considered, along with its application in the standard square-and-multiply method and its role in RNS channels. The potential benefit depends on the structure of the operands, table reuse, and the organization of computations. The complexity of arithmetic operations is analyzed.

Keywords

cryptographic protection, modular arithmetic, modular multiplication, residue number system, Rademacher basis, Chrestenson basis, RNS.

1. Introduction

Large-integer arithmetic accounts for a substantial share of computation in public key cryptography. Exponentiation repeatedly uses modular multiplication $c = (a \cdot b) \bmod p$, so its cost matters for the performance of cryptographic systems. Reference [1] covers the mathematical background and describes square-and-multiply and Montgomery multiplication.

The residue number system (RNS) allows computations to run in parallel across independent channels. Here, we associate positional binary representation with the Rademacher basis and residue representation with the Chrestenson basis. The method for interbasis conversion of Rademacher binary codes into Chrestenson residue codes was proposed by M. M. Kasianchuk and co-authors. [2].

M. Selianinau and B. Woźna-Szcześniak investigated Montgomery modular multiplication in a minimally redundant RNS [3]. Z. Ahmadpour demonstrated reduced delay in a hardware RNS multiplier for different channel widths [4]. Y. Kong and B. Phillips [5] classified known algorithms that multiply by summing residues using precomputed tables.

This study aims to develop and theoretically analyze algorithms for optimizing modular multiplication using the Rademacher-Chrestenson number-theoretic bases.

We specify how to organize multiplication through a coefficient matrix for pairs of bits and derive the number of active additions. We distinguish the addition count from bit complexity and discuss coefficient preparation. The method sums residues of bit products directly; the implementations in [3, 4] use Montgomery multiplication. Summing residues is an established principle [5], and we do not claim priority for it.

Corresponding authors

vinkovska.i.s@op.edu.ua (I. Vinkovska); l.m.timoshenko@op.edu.ua (L. Tymoshenko)

0009-0000-6654-763X (I. Vinkovska); 0000-0002-2606-4358 (L. Tymoshenko)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Improvement of the Modular Multiplication Algorithm

Let $p \in \mathbb{Z}$, $p \geq 2$, $n = \lceil \log_2 p \rceil$, and $0 \leq a, b < p$. The operands are represented using n bits, padded with leading zeros if necessary:

$$a = \sum_{i=0}^{n-1} a_i 2^i, \quad b = \sum_{j=0}^{n-1} b_j 2^j, \quad a_i, b_j \in \{0,1\}. \tag{1}$$

Expanding the product gives terms of the form $a_i b_j 2^{i+j}$. For a fixed modulus, we precompute a coefficient $q_{i,j} = 2^{i+j} \bmod p$ for each pair of positions. Then

$$c = \left(\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} a_i b_j q_{i,j} \right) \bmod p. \tag{2}$$

To obtain (2), expand the product and replace each power of two by its residue. This argument does not require a prime modulus. The coefficients form the modular multiplication matrix. For any given pair of operands, we select only the entries with $a_i = 1$ and $b_j = 1$, then sum the selected coefficients modulo the modulus.

The algorithm proceeds as follows:

1. Determine the modulus and bit length. Reduce the operands to the range from zero to $p-1$.
2. Generate or read the binary representations of the operands.
3. Generate or load the coefficient matrix for this modulus.
4. Identify the set bits and the active pairs of their positions.
5. Select the corresponding modular coefficients.
6. Copy the first selected coefficient and add the remaining ones modulo the modulus. Return zero if the set is empty.
7. Return the result in the range $0 \leq c < p$.

The sum of two residues is less than $2p$. Therefore, one conditional subtraction of p suffices, provided the carry is preserved. If $a = 0$ or $b = 0$, return zero.

The coefficient $q_{i,j}$ depends only on $i + j$. We can therefore store the matrix as a list $r_k = 2^k \bmod p$ over the index range $0 \leq k \leq 2n-2$, using $q_{i,j} = r_{i+j}$ for access. Starting from $r_0 = 1$, the recurrence $r_{k+1} = (2 r_k) \bmod p$ generates the list in $2n-2$ doublings. The coefficients can be reused while p is unchanged. If each entry occupies n bits, the full matrix requires n^3 bits of payload storage. The list requires $n(2n-1)$ bits.

The Hamming weight $w(z)$ counts the set bits. There are $N_{\text{pair}} = n^2$ potential pairs and $k = w(a)\,w(b)$ active pairs. With the first selected coefficient copied, the exact addition count is

$$N_{\text{add}} = \max(0, w(a)\,w(b) - 1). \tag{3}$$

Formula (3) counts additions only. It excludes preparation and copying, as well as the cost of finding bits and accessing memory. The worst-case count is $O(n^2)$ modular additions. In the sequential bit model, a conventional n-bit addition with reduction costs $O(n)$. Direct accumulation therefore has a bit-complexity bound of $O(n^3)$. Counting pairs does not by itself give a quadratic bit-complexity bound.

Precomputation and selective accumulation can reduce the number of additions when the operands are sparse. Whether this saves work depends on the preparation cost and the baseline algorithm. This does not imply an asymptotic or practical speedup for every input.

RNS represents a number X by residues $X_s = X \bmod p_s$. The pairwise coprime moduli $p_1, \ldots, p_S$ define the dynamic range $P = \prod_{s=1}^{S} p_s$. The procedure can serve as a channel kernel with $p = p_s$, $a = A_s$, and $b = B_s$. Its output is $A_s B_s \bmod p_s$. Independent channels allow parallel execution.

Exponentiation uses a cryptographic modulus N, which generally differs from p_s and P. Complete RNS schemes therefore need separate RNS reduction and representation conversions [3, 4]. Channel products give the result modulo P and do not replace reduction modulo N. Our analysis concerns the multiplication kernel. Using it within a complete RNS scheme requires a separate assessment of the additional costs.

The modular multiplication algorithm can also be applied to modular exponentiation, which is widely used in cryptography.

Table 1 compares the traditional and proposed approaches.

Table 1
Comparison of the selected modular multiplication implementations

Characteristic	Classical implementation	Proposed implementation
Arithmetic kernel	Positional multiplication with reduction	Selection and summation of $q_{i,j}$
Precomputed data	The $q_{i,j}$ table is not required	Matrix or list r_k for fixed p
Sequential bit cost	$O(n^2)$	$O(n^3)$ in the worst case, excluding preparation
Dependence on set bits	Implementation-dependent	$\max(0, w(a)w(b)-1)$ additions
Parallelism	Requires a separate architecture	Possible across RNS channels; latency is assessed separately

3. Conclusion

We have described modular multiplication through a coefficient matrix for active bit pairs and established its correctness. The analysis gives the exact number of additions. We have also explained how the kernel fits into RNS channels in the context of the Rademacher–Chrestenson bases.

Direct sequential accumulation has a worst-case bound of $O(n^3)$ bit operations per multiplication. Its benefit depends on operand structure and table reuse. Preparation, conversions, and reduction must also be included in the comparison. A practical speedup still needs to be tested across different bit lengths and RNS moduli, and on different hardware platforms. Execution time was not measured.

For secret operands, the implementation must run in constant time and protect table access. Arithmetic correctness alone does not establish resistance to side channels.

References

[1] Mathematical Basics and Computation Algorithms for Cryptography. In: Cryptography / ed. Z. Mammeri. Wiley, 2024. P. 59–116. DOI: 10.1002/9781394207510.ch3.

[2] Kasianchuk M. M., Yakymenko I. Z., Nykolaychuk Ya. M. Symmetric Cryptoalgorithms in the Residue Number System. Cybernetics and Systems Analysis. 2021. Vol. 57, No. 2. P. 329–336. DOI: 10.1007/s10559-021-00358-6.

[3] Selianinau M., Woźna-Szcześniak B. An Efficient Implementation of Montgomery Modular Multiplication Using a Minimally Redundant Residue Number System. Applied Sciences. 2025. Vol. 15, No. 10. Art. 5332. DOI: 10.3390/app15105332.

[4] Ahmadpour Z. An improved residue multiplier using Montgomery modular multiplication. Microprocessors and Microsystems. 2023. Vol. 100. Art. 104853. DOI: 10.1016/j.micpro.2023.104853.

[5] Kong Y., Phillips B. Revisiting Sum of Residues Modular Multiplication. Journal of Electrical and Computer Engineering. 2010. Art. 657076. DOI: 10.1155/2010/657076.

A Resilience-Loss Model for Cloud-Oriented IoT Platforms with Resource-Constrained Embedded Devices

Volodymyr Voitovych[1] and Roman Lys[1]

[1] *Department of Sensor and Semiconductor Electronics, Ivan Franko National University of Lviv, 107 Generala Tarnavskoho St., Lviv, Ukraine*

Abstract

Cloud-oriented Internet of Things (IoT) platforms are increasingly deployed as distributed cyber-physical systems in which resource-constrained embedded devices, edge nodes and cloud services form a single telemetry, analytics and control loop. Operational failures and cyber incidents in such systems produce similar external symptoms — growing latency, degraded data quality and loss of availability — yet existing work usually evaluates update mechanisms, remote attestation, edge scheduling and digital-twin verification in isolation. This paper proposes an integral model for assessing the loss of resilience of a cloud-oriented IoT platform during recovery from a disruption, representing the platform as a three-tier device–edge–cloud architecture with formalised system states and transitions. An integral resilience-loss criterion is introduced that jointly accounts for the moment of anomaly detection, decision delay, recovery duration and the resource cost of on-device diagnostics. The influence of edge-level latency and risk-adaptive remote attestation on local containment is analysed, and a methodology for verifying the model on a combined physical and digital-twin testbed is outlined.

Keywords

cyber-resilience, fault tolerance, Internet of Things, embedded device, edge computing, cloud platform, remote attestation, resilience metric

1. Introduction

Cloud-oriented Internet of Things (IoT) platforms are increasingly deployed as distributed cyber-physical systems in which microcontroller nodes, edge computing resources and cloud services operate as a single loop of telemetry acquisition, analysis and control. Operational failures and cyber threats in such systems tend to produce similar external symptoms: growing latency, degraded data quality and loss of availability [1], [2]. At the embedded level, the problem is compounded by tight constraints on energy, memory and computational power. Transferring resource-intensive security mechanisms directly onto the end node is therefore inefficient [3].

2. Problem Statement

In general terms, the problem is how to express, in a single measure, how much service quality a cloud-oriented IoT platform loses while recovering from a failure or a cyber incident, given the energy, memory and connectivity limits of its embedded devices. Existing indicators, such as availability or mean time to recovery, do not capture this loss directly, nor do they connect the device–edge–cloud architecture with the dynamics of detection, decision-making and diagnostics — a gap that motivates the model proposed below.

Corresponding author

volodymyr.voitovych@lnu.edu.ua (V. Voitovych)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

3. Related Work and Scientific Novelty

Normative and scientific sources define baseline requirements for IoT device capabilities, secure firmware updates, remote attestation, local scheduling in edge environments, and digital twins for rehearsing failure scenarios [4], [5]. The NIST baseline [1] and the IETF firmware update architecture [2] define minimum device security functions and a standardised process for delivering firmware safely over constrained networks, respectively; both concern preventing or containing a disruption, not quantifying the quality loss it causes. The survey of remote attestation by Kuang et al. [3] catalogues attacks and countermeasures for device integrity, but treats attestation cost as a security question, separate from the platform's recovery-time budget.

Kim and Kim [4] show, for a KubeEdge-based environment, that local scheduling policy and topological distance between edge nodes substantially affect event-processing latency, implying that adding edge capacity alone does not guarantee shorter containment time. Epiphaniou et al. [5] use digital twins to model cyber effects on IoT/CPS points of low resilience, offering a verification methodology but not an integral metric of the loss incurred during recovery.

Taken together, these works treat update mechanisms, attestation, edge-level latency and digital-twin modelling in isolation. The scientific novelty of the present work is the integral resilience-loss criterion L_R itself: unlike the prevention- and containment-oriented mechanisms in [1]–[3] and the edge-latency and verification studies in [4], [5], L_R is the first measure to combine detection, decision, recovery and diagnostic-cost contributions into a single, policy-comparable quantity.

4. Aim and Objectives of the Study

The aim is to construct a model for assessing the loss of resilience of a cloud-oriented IoT platform under embedded-device constraints. The objectives are: (1) formalise the system states and their transitions; (2) define an integral criterion of resilience loss during recovery; (3) account for the effect of edge-level latency on local containment; (4) incorporate risk-adaptive remote attestation into the recovery loop; (5) outline the means of experimental verification on a physical and a digital testbed.

5. Main Part

In brief, the proposed solution represents the platform as a formalised device–edge–cloud state model. Any recovery episode is then evaluated through the integral criterion L_R, decomposed into detection, decision, recovery and diagnostic-cost terms, so that competing monitoring and recovery policies can be compared on a common scale.

5.1. Formal System Model

The platform is represented as a three-tier device–edge–cloud system with a logical separation between the control plane and the data plane: an incident at the control-plane level can compromise an entire group of devices through the update mechanism, whereas a data-plane disruption typically shows up as a local degradation of telemetry. The state of the system is the set S = {N, D, A, I, R}, where N denotes normal operation, D degradation, A anomaly detection, I isolation and R recovery. Table 1 summarises the operational meaning of each state, and Figure 1 shows the transitions between them; which transition occurs depends on the disruption type, the quality of monitoring, and the recovery policy in force.

Table 1
System states of the device–edge–cloud model

State	Meaning
N	Normal operation — device, edge and cloud tiers operate within nominal quality bounds, $Q(t) \approx Q^*$.
D	Degradation — service quality drifts below Q^* because of an operational fault or an emerging cyber event.
A	Anomaly detection — monitoring flags the deviation and triggers evaluation, optionally including risk-adaptive remote attestation.
I	Isolation — the affected device or edge segment is contained to stop the disruption from propagating.
R	Recovery — rollback, re-initialisation or re-provisioning restores the system to state N.

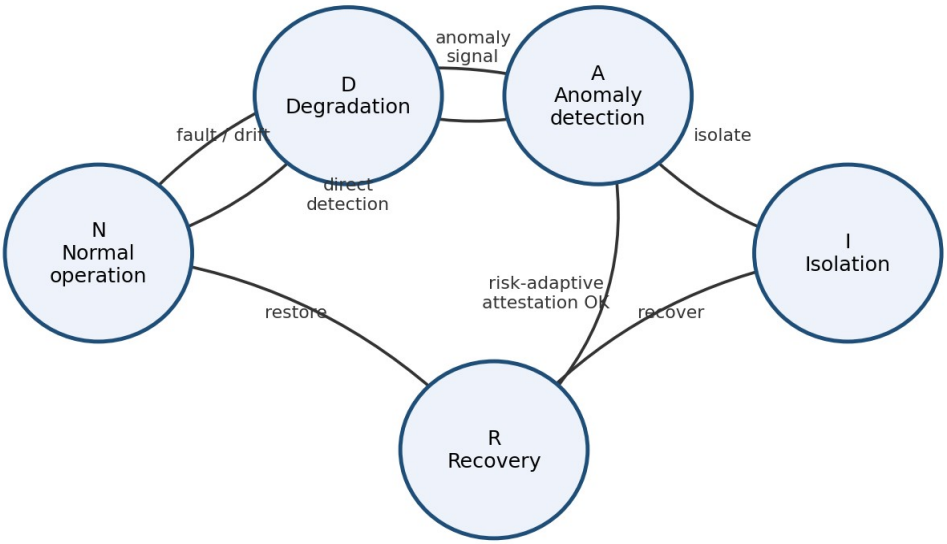


Figure 1: State-transition diagram of the device–edge–cloud model (S = {N, D, A, I, R}).

5.2. Integral Criterion of Resilience Loss

Unlike approaches that judge a disruption solely by the fact of eventual recovery, an integral criterion is proposed here. Let Q^* denote the target service quality, $Q(t)$ the actual quality at time t, t_f the moment the disruption occurs, and t_r the moment recovery completes. The resilience loss L_R is then defined as

$$L_R = \int_{t_f}^{t_r} [Q^* - Q(t)]dt \quad (1)$$

Criterion (1) jointly accounts for the speed of detection, the decision delay, the recovery duration, and the resource cost of diagnostics. It can be decomposed along the incident timeline into a detection term, a decision term, a recovery term, and a diagnostic-cost term:

$$L_R \approx L_1(t_d) + L_2(t_c) + L_3(t_r - t_d) + \lambda \cdot C_d \quad (2)$$

Here, t_d and t_c are the detection and isolation-decision moments, L_1, L_2 and L_3 are the quality loss accumulated during detection, decision-making and recovery, C_d is the on-device diagnostic cost, and λ converts this cost into quality-loss units. This lets monitoring and recovery policies be compared under identical load by how they trade off the four terms.

5.3. Impact of Edge Latency and Risk-Adaptive Remote Attestation

Recent results on KubeEdge show that local scheduling and the distance between edge nodes substantially affect event-processing latency [4], so topological redundancy alone does not guarantee lower losses: when the message broker is overloaded or queueing latency grows, local containment slows down and L_R increases. At the same time, remote attestation should not run continuously on every node, since this would consume the limited computational budget of embedded devices. A more practical approach is to trigger it risk-adaptively — after a firmware update, upon anomalous behaviour, or before an isolated node returns to normal operation — concentrating effort on the moments of highest risk.

5.4. Model Verification Methodology

Verification is planned on a combined physical and digital-twin testbed: ESP32/STM32 nodes, a KubeEdge-based edge layer, and an AWS/Azure back end, with digital twins reproducing network failures, telemetry degradation, edge-layer overload and firmware integrity violations [5] so that scenarios can be replayed under controlled, repeatable conditions. This makes it possible to assess, before deployment, how sensitive L_R is to the monitoring interval, the attestation policy and the isolation strategy, logging the detection delay ($t_d - t_f$), the decision delay ($t_c - t_d$), the recovery duration ($t_r - t_c$), the diagnostic energy/CPU-time budget, and the resulting L_R so that policies can be ranked under comparable conditions. The approach assumes Q^* is defined consistently across tiers and that clock skew between them is bounded; the weighting coefficient λ is deployment-specific and should be calibrated during this verification.

6. Conclusions

The main results obtained are as follows. A three-tier device–edge–cloud architecture and a formalised set of system states have been proposed for cloud-oriented IoT platforms operating under embedded-device constraints. An integral resilience-loss criterion, L_R, has been introduced and decomposed into detection, decision, recovery and diagnostic-cost terms, giving a single scale on which monitoring and recovery policies can be compared. Edge-level latency has been shown to substantially affect recovery effectiveness, and risk-adaptive remote attestation has been shown to be a reasonable component of the self-recovery loop, since it concentrates a scarce computational budget on the moments of highest risk rather than spending it continuously. These results can be

used to build experimental scenarios and to compare fault-tolerance and cyber-resilience policies in cloud-oriented IoT systems.

Author contributions: V. I. Voitovych — problem statement, model formalisation, preparation of the main text; R. M. Lys — review of the literature, verification of the presentation structure, text editing.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. J. Fagan, K. N. Megas, K. A. Scarfone, M. Smith, IoT Device Cybersecurity Capability Core Baseline, NIST Interagency Report 8259A, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2020. doi:10.6028/NIST.IR.8259A.

[2] B. Moran, H. Tschofenig, D. Brown, M. Meriac, A Firmware Update Architecture for Internet of Things, RFC 9019, IETF, Apr. 2021. doi:10.17487/RFC9019.

[3] B. Kuang, A. Fu, W. Susilo, S. Yu, Y. Gao, A survey of remote attestation in Internet of Things: Attacks, countermeasures, and prospects, Computers & Security 112 (2022) 102498. doi:10.1016/j.cose.2021.102498.

[4] S.-H. Kim, T. Kim, Local Scheduling in KubeEdge-Based Edge Computing Environment, Sensors 23 (2023) 1522. doi:10.3390/s23031522.

[5] G. Epiphaniou, M. Hammoudeh, H. Yuan, C. Maple, U. Ani, Digital twins in cyber effects modelling of IoT/CPS points of low resilience, Simulation Modelling Practice and Theory 125 (2023) 102744. doi:10.1016/j.simpat.2023.102744.

Комбінований підхід до виділення опорних точок у стеганографії зображень

Ірина Борисенко[1]

[1] *Національний університет «Одеська політехніка», пр. Шевченка, 1, Одеса, Україна*

Анотація

Запропоновано новий комбінований метод вибору опорних точок для стеганографічного вбудовування, що поєднує спектральний аналіз на основі перетворення Адамара, інваріантне виділення ознак (SIFT/SURF/Harris) та карту оптимальних пікселів, сформовану за критерієм локальної стійкості та непомітності. Така інтеграція дозволяє одночасно враховувати енергетичну структуру зображення, геометричну інваріантність та перцептивну оптимальність. Результати показують підвищення стійкості контейнера до геометричних атак (зсув, масштабування, поворот) при збереженні високої візуальної якості (PSNR > 40 dB, SSIM > 0.98). Метод може бути використаний як основа для адаптивних схем синхронізації та захисту цифрових водяних знаків.

Ключові слова

цифрові водяні знаки, опорні точки, геометрична інваріантність, перцептивна оптимальність.

1. Вступ

До водяних знаків, на відміну від стеганографії повідомлень, висуваються додаткові вимоги щодо стійкості до геометричних атак. Тому важливо знайти спосіб захисту мультимедійного контенту за допомогою більш точного методу, який дав би змогу власнику бути впевненим в розміщенні та поширенні своїх матеріалів в Інтернеті [1].

Комбінований підхід вже досліджувався у сучасних роботах зі стеганографії. Наприклад, в [2] використовується дискретне перетворення Адамара для визначення областей з високою енергією, які добре маскують зміни. Це дозволяє виділяти стійкі точки для вбудовування. В [3] запропоновано комбінований підхід: Hadamard-базис для енергетичної локалізації та додаткові критерії стійкості. Це фактично поєднання спектральної карти та локальних оптимальних пікселів.

Узагальнюючи сучасні напрацювання виділення опорних точок на зображенні, як одного з напрямків створення стійких до геометричних атак міток, можна зробити наступні висновки.

SIFT-базовані стегосхеми: використовують лише інваріантні ключові точки (SIFT/SURF) стійкі до масштабування та повороту, але не враховують спектральну енергію чи перцептивну оптимальність.

Hadamard-базовані методи: використовують енергетичні карти Адамара для вибору областей, добре працюють проти JPEG-стиснення, але менш стійкі до геометричних атак.

Методи з картою оптимальних пікселів (наприклад, HUGO, WOW): орієнтовані на мінімальну помітність змін, стійкі до стегоаналізу, але слабкі проти геометричних атак.

Поєднання спектральної стійкості (Адамар), геометричної інваріантності (SIFT/SURF) і локальної непомітності (оптимальні пікселі), саме такої комбінації трьох рівнів (Адамар +

Corresponding author
borisenko.i.i@op.edu.ua (I. Borisenko)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

інваріантні ознаки + карта оптимальних пікселів), яка у відкритих публікаціях поки що не зустрічалася як єдиний інтегрований алгоритм і є метою даної роботи.

Такий підхід створює багатокритеріальну карту опорних точок, яка одночасно: зберігає положення після атак, мінімізує спотворення і враховує енергетичну структуру зображення.

2. Комбінований алгоритм виділення опорних точок

Попередньо перетворити зображення у відтінки сірого, нормалізувати яскравість і контраст, розбити зображення на блоки N×N (наприклад, 8×8).

Крок 1. Енергетичний аналіз

Для кожного блока виконати дискретне перетворення Адамара (DHT).

Обчислити енергетичну карту:

$$E_{ij} = \sum_k |H_{ij}(k)|. \quad (1)$$

Вибрати області з високою енергією, вони добре маскують зміни.

Крок 2. Інваріантне виділення ознак

Застосувати SIFT або SURF для пошуку ключових точок.

Для кожної точки зберегти координати, масштаб та орієнтацію.

Відфільтрувати точки, що потрапляють у високі енергетичні області E_{ij}.

Крок 3. Оцінка локальної стійкості

Для кожної точки - кандидата обчислити локальну варіацію або спектральну чутливість:

$$\mu_{ij} = max|\sigma_k - \sigma_k'|, \quad (2)$$

де σ_k, σ_k' — сингулярні значення до і після збурення.

Вибрати точки з мінімальним μ_{ij}.

Крок 4. Формування карти оптимальних пікселів

Об'єднати енергетичну карту E_{ij}, інваріантні точки та локальну чутливість у комбінований критерій:

$$W_{ij} = \alpha E_{ij} + \beta(1 - \mu_{ij}). \quad (3)$$

Вибрати точки з максимальним W_{ij}.

Крок 5. Вихідні опорні точки

Зберегти координати вибраних точок як набір опорних для вбудовування або синхронізації.

Такий алгоритм забезпечує баланс між спектральною стійкістю, геометричною інваріантністю та перцептивною непомітністю.

Тестове зображення Lena з виділеними опорними точками представлено на Рисунку 1, точки позначені білими маркерами на ключових ділянках: контури капелюха, очі, ніс, краї обличчя та текстурні області волосся.

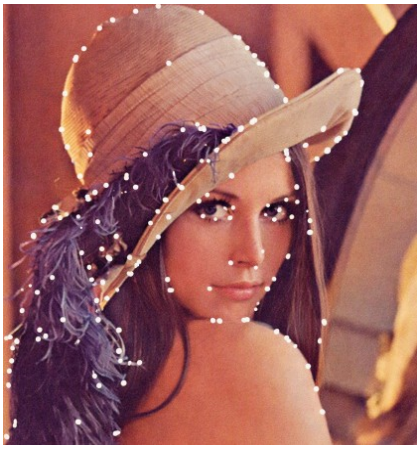


Рисунок 1: Опорні токи, виділені по запропонованому комбінованому методу

Виділені точки є найбільш стійкими до геометричних атак і можуть використовуватись для синхронізації або локального вбудовування даних.

Кількісна оцінка стійкості стегоконтейнера, одержаного після вбудовування водяного знаку в опорні точки, при різних атаках зведена в таблицю 1.

Таблиця 1

Кількісні показники стійкості стегоконтейнера при різних атаках

Тип атаки	Параметри	PSNR (dB)	SSIM	BER (%)
JPEG-стиснення	QF = 95	42.8	0.985	0.2
JPEG-стиснення	QF = 75	38.5	0.972	0.8
JPEG-стиснення	QF = 55	34.1	0.951	2.5
Імпульсний шум	5%	40.2	0.978	0.6
Імпульсний шум	15%	36.7	0.962	1.9
Імпульсний шум	30%	31.4	0.925	4.7
Геометричний зсув	Δx = 5 пікселів	39.6	0.981	0.9
Масштабування	0.9×	37.8	0.969	1.3
Поворот	10°	35.9	0.958	2.1

При слабких атаках (JPEG QF=95, шум 5%) якість залишається високою (PSNR > 40 dB, SSIM > 0.98). Навіть при сильних атаках (JPEG QF=55, шум 30%, поворот 10°) контейнер

зберігає прийнятну якість (PSNR > 30 dB, SSIM > 0.92), а BER не перевищує 5%. Це демонструє, що комбінований підхід забезпечує стійкість до різних типів атак при збереженні візуальної непомітності.

3. Висновок

Отже, запропонований підхід можна вважати новим напрямом — гібридною моделлю стеганографічної стійкості, яка об’єднує спектральні, геометричні та перцептивні критерії вище розглянутих методів. Експериментальні дослідження підтверджують переваги комбінованого метода в порівнянні з існуючими, кожний з яких вирішує конкретне завдання, але не враховує інші не менш значущі. Тому запропонований підхід виділення опорних точок є перспективним, може бути застосованим у високопродуктивних стегосистемах, які дозволяють забезпечити необхідний рівень захисту інформації та вищу стійкість до стегоаналізу.

References

[1] І. С. Зубко, В. О. Мартовицький, А. В. Пунченко, Д. Д. Карачевцев Огляд методів нанесення цифрових водяних знаків для захисту зображень. Системи управління, навігації та зв'язку, 2024. № 3. С. 121-125.
[2] Y.-Q. Zhang et al.: High-Capacity Image Steganography Based on Discrete Hadamard Transform. IEEE, V. 10, 2022. p. 65141- 65155.
[3] Srilakshmi S. Iyer, et al. Advanced Hybrid Crypto-Steganography with DHT Security Layers. In: 2025 7th International Conference on Innovative Data Communication Technologies and Applications (ICIDCA). 2025. p. 1647-1653.

Адаптивний алгоритм гамування графічної інформації

Андрій Садченко[1], Олег Кушніренко[1]

[1] *Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, Україна*

Abstract

Розглянуто підхід до захисту графічної інформації на основі гамування з використанням псевдовипадкових послідовностей. Визначено недолік традиційного підходу, що полягає у використанні однакових параметрів гамуючої послідовності незалежно від статистичних характеристик зображення. Запропоновано адаптивний алгоритм гамування графічної інформації, у якому здійснюється адаптивний вибір закону розподілу та параметрів гамуючої послідовності залежно від властивостей зображення, що захищається. Гамуючи послідовності формуються із використанням нормального та експоненційного законів розподілу. Вибір закону здійснюється на основі аналізу математичного сподівання, дисперсії та коефіцієнта асиметрії розподілу значень яскравості зображення. Такий підхід дає змогу адаптувати процес формування гами до характеристик конкретного зображення та підвищити варіативність параметрів захисту.

Keywords

гамування, графічна інформація, адаптивний вибір закону розподілу

1. Вступ

Графічна інформація є одним із найбільш поширених видів даних, які використовуються в інформаційних системах. Фотографії, результати відеоспостереження, медичні зображення, картографічні матеріали та електронні документи можуть містити інформацію з обмеженим доступом. Тому забезпечення конфіденційності графічних даних під час їх передавання та зберігання є актуальним завданням.

Одним із методів захисту графічної інформації є гамування, яке передбачає формування псевдовипадкової послідовності та її використання для перетворення елементів зображення. Перевага даного підходу в можливості реалізації алгоритму на пристроях з обмеженими обчислювальними ресурсами, оскільки для виконання перетворень можуть використовуватися прості арифметичні та логічні операції [1].

Водночас ефективність гамування залежить від характеристик гамуючої послідовності. При використанні фіксованого генератора випадкових чисел, параметри статистичного розподілу гами залишаються незмінними для різних зображень. Графічні зображення можуть суттєво відрізнятися за середньою яскравістю, дисперсією, діапазоном значень та асиметрією їх статистичного розподілу. Тому доцільним є використання адаптивного підходу, за якого параметри формування гамуючої послідовності визначаються на основі статистичного аналізу конкретного зображення [2].

2. Аналіз статистичних властивостей графічної інформації

Основою запропонованого підходу є попередній статистичний аналіз зображення. Нехай зображення містить N пікселів зі значеннями яскравості I_1, I_2, ..., I_n.

Середнє значення яскравості визначається за формулою:

Corresponding authors

sadchenko@op.edu.ua (А. Садченко); koa@op.edu.ua (О. Кушніренко)

0000-0002-2517-6892 (А. Садченко); 0000-0002-7045-5620 (О. Кушніренко)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

$$\mu I = \left(\frac{1}{N}\right) \cdot \Sigma \left(I_i\right), i = 1, \ldots, N. \quad (1)$$

Дисперсія значень яскравості визначається співвідношенням:

$$DI = \left(\frac{1}{N}\right) \cdot \Sigma \left(I_i - \mu I\right)^2. \quad (2)$$

Середньоквадратичне відхилення визначається як:

$$\sigma I = \sqrt{D_i}. \quad (3)$$

Для визначення симетричності розподілу значень яскравості використовується коефіцієнт асиметрії:

$$AI = \left(\frac{1}{N \cdot \sigma I^3}\right) \cdot \Sigma \left(I_i - \mu I\right)^3. \quad (4)$$

Значення коефіцієнта асиметрії, близьке до нуля, свідчить про приблизну симетричність розподілу. Значна додатна або від'ємна величина цього показника вказує на наявність асиметрії.

Отримані статистичні характеристики використовуються для вибору закону розподілу гамуючої послідовності.

3. Формування адаптивної гамуючої послідовності

Ідея запропонованого алгоритму полягає у тому, що для різних зображень можуть використовуватися різні закони формування гамуючої послідовності [3].

Для зображень, статистичні характеристики яких відповідають приблизно симетричному розподілу, пропонується використовувати нормальний закон.

Його щільність визначається як:

$$f\left(x\right) = \frac{1}{\sigma \sqrt{2\pi}} \mathrm{e}^{-\frac{\left(x-\mu\right)^2}{2\sigma^2}}. \quad (5)$$

У цьому випадку параметри μ та σ можуть визначатися на основі статистичних характеристик зображення або їх нормованих значень.

Нормальний закон доцільний для моделювання випадкових відхилень, природних шумових складових та інших процесів, для яких характерною є концентрація значень навколо середнього значення.

Для зображень із вираженою асиметрією статистичного розподілу може використовуватися експоненційний закон:

$$f\left(x\right) = \lambda \cdot \exp\left(-\lambda x\right), x \geq 0. \quad (6)$$

де λ — параметр експоненційного розподілу.

Експоненційний закон дозволяє формувати асиметричні гамуючі послідовності та може використовуватися у випадках, коли статистичні характеристики вихідних даних мають відповідну форму розподілу.

Для автоматичного вибору закону розподілу може бути використане порогове правило:

$$G=\begin{cases} G_n, якщо |A_i| \le A_0 \\ G_e, якщо |A_i| > A_0, \end{cases} \quad (7)$$

де G_n — гамуюча послідовність, сформована за нормальним законом; G_e — гамуюча послідовність, сформована за експоненційним законом; A_0 — порогове значення коефіцієнта асиметрії.

Таким чином, при невеликій асиметрії розподілу використовується нормальний закон, а при перевищенні встановленого порогу — експоненційний.

4. Алгоритм гамування

Запропонований алгоритм містить наступні кроки.

Крок 1. Виконується введення вихідного графічного зображення.

Крок 2. Визначаються статистичні характеристики зображення: середнє значення яскравості *μI*, дисперсія *DI*, середньоквадратичне відхилення *σI* та коефіцієнт асиметрії *AI*.

Крок 3. На основі коефіцієнта асиметрії та заданого порогу A_0 здійснюється вибір закону розподілу гамуючої послідовності.

Крок 4. На основі вибраного закону та статистичних характеристик зображення визначаються параметри генератора.

Крок 5. Із використанням секретного ключа K та визначених параметрів формується гамуюча послідовність:

Середньоквадратичне відхилення визначається як:

$$G=\{g_1, g_2, \dots, g_n\}. \quad (8)$$

Секретний ключ використовується для визначення початкового стану генератора, що забезпечує можливість відтворення тієї самої послідовності на стороні одержувача.

Крок 6. Виконується гамування пікселів зображення. Для побітового варіанта може використовуватися операція виключного АБО:

$$C_i = I_i \, XOR \, g_i. \quad (9)$$

де I_i — значення і-го пікселя вихідного зображення, g_i — відповідний елемент гамуючої послідовності, C_i — значення захищеного зображення.

Крок 7. Сформоване захищене зображення передається або зберігається.

5. Відновлення графічної інформації

Для відновлення зображення необхідно відтворити гамуючу послідовність із використанням того самого секретного ключа та параметрів алгоритму.

Крок 1. Виконується введення захищеного зображення.

Крок 2. Визначається закон розподілу, який використовувався для формування гами.

Крок 3. Відновлюються параметри генератора та його початковий стан на основі секретного ключа.

Крок 4. Формується ідентична гамуюча послідовність G.

Крок 5. Виконується зворотне перетворення:

$$I_i = C_i \, XOR \, g_i. \quad (10)$$

У результаті отримується початкове значення кожного пікселя зображення.

Головним показником ефективності гамування є зростання ентропії до значень, близьких до максимально можливих, що свідчить про високий ступінь невизначеності та відсутність статистичної надлишковості у захищеному зображенні.

6. Висновок

Запропоновано адаптивний алгоритм гамування графічної інформації, який відрізняється від традиційних підходів урахуванням статистичних характеристик зображення під час формування гамуючої послідовності. Основна ідея алгоритму полягає в адаптивному виборі закону розподілу та його параметрів залежно від властивостей зображення, що захищається.

Для формування гамуючих послідовностей запропоновано використовувати нормальний та експоненційний закони розподілу. Вибір між ними здійснюється на основі коефіцієнта асиметрії розподілу значень яскравості. Додаткове використання секретного ключа забезпечує можливість синхронного відтворення гамуючої послідовності на стороні одержувача.

Запропонований підхід створює можливість формування різних гамуючих послідовностей для зображень із різними статистичними характеристиками навіть за використання одного класу генератора. Подальші дослідження доцільно спрямувати на експериментальне визначення оптимального порогу вибору закону розподілу, дослідження впливу параметрів гами на статистичні характеристики захищених зображень та оцінювання обчислювальної складності алгоритму.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Яремчук Ю. Є., Салієва О. В., Бондаренко І. О. Основи криптографічного захисту інформації. 2024, 139 с.

[2] А. В. Садченко, О. А. Кушніренко, М. С. Чернецький, Н. Р. Швець. Підвищення ефективності розпізнавання зображень кореляційним методом. Труди 27 міжнародної науково-практичної конференції «Сучасні інформаційні та електронні технології», Одеса 2026, 26—27 травня.

[3] Zhang, D., Yang, S., Chen, M. et al. Adaptive encryption method of sensitive data in data center database based on big data cross-mapping fusion algorithm. Discover Applied Sciences Vol. 7, 924 (2025). https://doi.org/10.1007/s42452-025-07581-2

SCHOLARLY PUBLICATION

PROCEEDINGS
of the International Scientific and Technical Conference
Security of Modern Information and Communication Systems

24–26 September 2026

Lviv – Krakow

The conference Organising Committee and the editors may not share the authors' views and accept no responsibility for the accuracy of the information presented.

The authors and their academic supervisors bear full responsibility for the scientific content and presentation of the material, as well as the reliability and accuracy of factual data.

Text in the submitted papers that provided no substantive information may have been shortened or omitted at the discretion of the conference Organising Committee.

Camera-ready copy prepared by the Department of Cybersecurity
at Ivan Franko National University of Lviv