Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Explanations to Deployable Rules: A Systematic Review of Explainable AI (XAI)-Guided Rule Extraction in Network Intrusion Detection

Mandahadi Kusuma
University of Technology Malaysia
Mohd Yazid Idris
yazid@utm.my

University of Technology Malaysia
Adila Firdaus Arbain
University of Technology Malaysia
Mohd Kufaisal Mohd Sidik
University of Technology Malaysia

Systematic Review

Keywords: explainable AI, SHAP, rule extraction, network intrusion detection, deployment readiness, transferability, adaptive rule, DDoS

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11453291/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

From Explanations to Deployable Rules: A Systematic Review of Explainable AI (XAI)-Guided Rule Extraction in Network Intrusion Detection

Mandahadi Kusuma[a,1,2], **Mohd Yazid Idris**[b,1,2,3], **Adila Firdaus binti Arbain**[c,1,2], **Mohd Kufaisal bin Mohd Sidik**[d,1,4,2]

[1]Universiti Teknologi Malaysia, Johor Bahru, Johor 81310, Malaysia
[2]Faculty Of Computing
[3]Media and Game Center of Excellence(MaGICX), Institute of Human Centered Engineering(IhumEn)
[4]Department of Computer Science
Received: date / Accepted: date

Abstract This systematic review examines the intersection of explainable artificial intelligence (XAI) and deployable security rule extraction in network intrusion detection systems (NIDS). Following the PRISMA methodology, we analyzed 109 papers published between January 2022 and July 2026 using a four-category quality scoring system, a hierarchical taxonomy, statistical analysis, and thematic analysis. SHAP (SHapley Additive exPlanations) is the most widely adopted XAI method, used in 77 papers (70.6%), yet only 15 papers (13.8%) produce deployment-ready rules. We identified two research gaps. First, an accuracy–deployment disconnect caused by a rule translation gap: 58 papers report an accuracy of at least 99%, 51 of them non-deployable and 7 deployment-ready. However, this accuracy rarely becomes deployable rules. Rule extraction quality (Rule_Score) shows a significant association with deployment readiness (point-biserial $r = 0.455$, $p < 0.0001$) with a large effect size (Cohen's $d = 1.47$), whereas XAI technique quality (XAI_Score) shows no significant association with deployment readiness ($r = 0.138$, $p = 0.1533$). Second, a transferability gap: 33 papers report no transferability validation, 55 papers validate along a single dimension, and only 5 papers validate across all three dimensions (dataset, network environment, and attack type). Furthermore, only one paper (0.9%) includes an adaptive update mechanism that updates deployable rules as network conditions change. This review establishes new research opportunities: (1) automated rule extraction pipelines that transform XAI outputs into deployable formats, (2) adaptive rule frameworks that update rules as network conditions evolve, and (3) evaluation standards that combine deployment-readiness assessment and transferability across all three dimensions.

Keywords: explainable AI, SHAP, rule extraction, network intrusion detection, deployment readiness, transferability, adaptive rule, DDoS

1 Introduction

1.1 Problem Background

Current machine learning (ML) and deep learning (DL) models achieve detection accuracy of 99% or higher on benchmark intrusion detection datasets [1, 2]. Despite this accuracy, most ML/DL models operate as black boxes, providing little insight into their decision-making processes. This lack of transparency creates an opaque barrier for security analysts who need to understand, trust, and act upon the alerts these systems generate. Then the critical question arise whether these black-box models can be translated into interpretable, deployable, and adaptive security rules for real-world network environments.

Explainable Artificial Intelligence (XAI) techniques have emerged as a promising solution to this interpretability challenge. XAI methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) provide insights into model predictions by highlighting feature importance and decision pathways [3, 4]. However, while these techniques can display model behavior, they often stop on features explanation only, without generating actionable security rules that can be directly implemented in network intrusion detection systems (NIDS).

[a]e-mail: mandahadi@graduate.utm.my
[b]e-mail: yazid@utm.my
[c]e-mail: adilafirdaus@utm.my
[d]e-mail: kufaisal@utm.my

1.2 Systematic Review Objectives

This systematic review examines the intersection of explainable artificial intelligence, network security, and operational deployment. The current state and future direction of XAI-guided rule extraction for network intrusion detection are explored through four objectives:

- **Objective 1 (Systematic Analysis of XAI Techniques).** To identify, categorize, and assess the XAI techniques employed in network intrusion detection research.
- **Objective 2 (Evaluation of Rule Extraction Methods).** To evaluate and categorize methods for extracting deployable rules from ML/DL models, from model explanation to operational rule generation, and to examine whether detection accuracy, XAI technique quality, or rule extraction quality distinguishes deployment-ready research.
- **Objective 3 (Assessment of Network Security Relevance).** To evaluate how XAI-guided rule extraction approaches address network intrusion detection requirements, with particular emphasis on DDoS attack detection.
- **Objective 4 (Identification of Methodological Gaps and Limitations).** To identify current research limitations in transferability, adaptive rule-update mechanisms, and deployment readiness.

1.3 Limitations of the research

This systematic review has several limitations:

- The focus on English-language publications, may miss important work in other languages.
- The review period (January 2022 – July 2026) captures the most recent research on XAI-based intrusion detection but excludes earlier foundational work that established concepts or methodologies still relevant to current research.
- The search scope focused explicitly on "XAI + Rules", which may miss implicit rule generation approaches that do not mention "rules" in their titles, abstracts, or keywords despite generating deployable rule sets under other terminology.
- Although a strict scoring system was applied, some subjectivity remains in categorizing papers, particularly when papers do not explicitly state their methodological limitations or deployment readiness.
- Papers that report only other metrics (such as F1-score, false-positive rate, or rule fidelity) are treated as not reporting accuracy. Detection accuracy value was the highest overall accuracy reported in each paper. However these values are not directly comparable, therefore they are summarized descriptively rather than compared inferentially.
- Explanation fidelity and computational efficiency, which are evaluated in [5, 6], are outside the scope of the scoring used here.

2 Related Works

2.1 Machine Learning Network Intrusion Detection Systems

The breakthrough in integrating machine learning into NIDS came with the application of deep learning techniques. Studies by [7, 8] demonstrate that deep learning models, particularly Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, can achieve detection rates exceeding 99% for various attack types, including DDoS attacks. Recent research on DDoS classification shows that a tuned Random Forest (RF) can reach a static accuracy of 99.99% on a DDoS-SDN dataset [9]. However, this high accuracy comes at a price: the models work as black boxes and provide little to no insight into their decision-making processes, and the thousands of rule splits in an RF make it far less interpretable. These drawbacks create an interpretability and deployment gap.

2.2 Current XAI Techniques and Their Limitations

Most popular XAI method is SHAP (SHapley Additive exPlanations). This method provides theoretically grounded feature importance explanations based on cooperative game theory[3]. LIME (Local Interpretable Model-agnostic Explanations) is less popular than SHAP, often complementing SHAP for local explanations. Research use LIME is CMIR-Gen automatic signature generation algorithm for IDS applications across multiple attack types[10].

Custom XAI methods, represent the most innovative approaches but also the most challenging to generalize. The interpretable generalization mechanism[2] introduces novel set-theoretic approaches to explanation, while the abnormal logical representation learning framework [11] creates inherently interpretable neural-symbolic architectures. These custom methods often achieve superior domain-specific performance but lack the generalizability and standardization necessary for widespread adoption.

Another approach are Attention mechanisms as primary XAI techniques with attention-augmented approaches in some research paper like [12]. Despite showing high accuracy potential at 97-99%, it still lacks of rule extraction component.

Current XAI applications in network intrusion detection still have several limitations:

- **Computational Overhead:** XAI computation, especially for large datasets, takes more longer time [13][14], lead-

ing to delays in real-time intrusion detection applications where rapid response is critical.
- **Limited Cross-Dataset Evaluation:** Many XAI evaluations are limited to a single dataset[15][16], hindering understanding of cross/domain generalizability and raising questions about whether explanations remain valid when applied to different network environments or attack scenarios.
- **Ensemble Complexity:** While ensemble ML models for IDS can increase detection accuracy, applying XAI to ensemble architectures introduces additional complexity layers that make it challenging to extract coherent, unified rule sets from multiple model outputs[17][18][4].

2.3 XAI Applications in Network Security

XAI has been applied to domain-specific network security, for example in explainable malicious traffic detection for industrial IoT networks [19], and hybrid IDS designs combine supervised and unsupervised learners [20] or rule-based Security Information and Event Management (SIEM) systems with AI-based detectors [21]. Most of this work, however, stops at feature-importance explanations: the XAI-IDS framework applies SHAP and LIME across seven black-box models [1], and an interpretable generalization mechanism introduces intersection-based interpretability for network traffic [2], but neither turns its explanations into deployable rules. Hybrid explainable NIDS frameworks recognize that explanations must translate into practical security operations [22], and reviews and evaluations of XAI methods for intrusion detection [23, 24] compare how well the explanations describe model decisions, yet how these explanations can be converted systematically into operational security rules remains open.

This needs also reinforced by NIST standards that intrusion detection systems (IDS) should provide detailed logging capabilities supporting incident handling and forensic analysis[25]. Not only simple detection flags, but requirement of interpretability throughout the detection pipeline also important.

2.4 Rule Translation in Intrusion Detection Systems

The translation of DL/ML model logic into human-readable DDoS rule filters is crucial for practical NIDS deployment. Various methods have been proposed for rule extraction, categorized as pedagogical (black-box), decompositional (white-box), or eclectic (hybrid) approaches[26]

Current research in rule generation is the Anomaly2Sign framework[27], which automatically transforms anomaly detection models into interpretable decision trees, enabling direct conversion from model predictions to IF-THEN rule formats suitable for deployment in standard intrusion detection systems. This represents a promising direction for automated rule extraction, though the framework's effectiveness for complex temporal attack patterns like sophisticated DDoS remains to be validated across diverse network environments.

Network environments and attack patterns evolve continuously, yet current rule generation approaches produce static rules that cannot adapt to these changes. Studies like[28][29] acknowledge there is performance degradation over time without offer any systematic solutions for rule adaptation. The absence of adaptive rule modification frameworks means that organizations must choose between outdated rules with degrading performance or run complex ways to update the rules.

2.5 Comparison with Existing Reviews

Several reviews published between 2021 and 2026 have examined rule-based or explainable intrusion detection, but they differ in scope, protocol, and depth of analysis. Table 1 compares this review with seven of them across eleven aspects relevant to deployable rule extraction. Three of the seven follow an explicit review protocol: Mohale and Obagbuwa [32] apply PRISMA to 20 studies on XAI in IDS published between 2017 and 2023; Fasla and Mehmood [34] apply PRISMA to 81 studies on web application threat detection and explicitly exclude network-level intrusion detection without payload analysis; and Souto et al. [5] follow the Kitchenham and Charters guidelines with snowballing to analyse 71 studies on explainable intrusion detection for IoT. The remaining four [6, 30, 31, 33] are narrative or comparative reviews.

This presented systematic review studies have four differences. First, this review statistically tests whether the quality dimensions of each included paper are associated with deployment readiness, whereas the compared reviews report descriptive summaries only. Second, none of the compared reviews classifies deployment readiness for each reviewed paper, whereas this review classifies every included paper. Third, none analyses transferability along separate dimensions, whereas this review codes dataset, network-environment, and attack-type validation for each included paper. Fourth, although adaptive rule maintenance is discussed conceptually by Sarker et al. [31] and Liu et al. [30], none of the compared reviews examines how often the reviewed papers implement it, whereas this review records whether each included paper provides an adaptive rule-update mechanism. In addition, none of the compared reviews treats DDoS detection as a distinct focus, whereas this review reports DDoS-specific papers separately.

These differences should be read within the limits of the comparison. This review is not the first to use a scored quality rubric [5], nor the first to identify adaptive rule maintenance

Table 1 Comparison of this review with existing reviews on rule-based and explainable intrusion detection

Reviews \ Aspects	Years Covered	Number of Studies	Systematic Review Protocol	Scored Quality Assessment	Inferential Statistics	Rule Extraction Analysis	Rule or Explanation Quality Criteria	Deployment Readiness	Transferability Dimensions (0–3)	Adaptive Rule Maintenance	DDoS-Specific Analysis	XAI Technique Taxonomy	Evidence-Derived Requirements	Remarks
Liu et al. (2021) [30]	1998–2020[a]	21[a]	×	×	×	✓	✓	△	0	△	×	×	△	Narrative survey of rule learning-based IDS with smart-grid prospects; proposes six rule-quality criteria
Sarker et al. (2024) [31]	2010–2023	130[b]	△	×	×	✓	△	×	0	△	×	×	△	Narrative review of rule-based AI for critical infrastructure security; taxonomy of rule generation
Mohale & Obagbuwa (2025) [32]	2017–2023	20	△	△	×	△	△	×	0	×	×	△	△	PRISMA review of XAI in IDS; thematic synthesis of SHAP, LIME, and rule- or tree-based models
Khediri et al. (2025) [33]	2019–2025	n.s.	×	×	×	△	△	×	0	×	×	✓	△	Narrative survey proposing a four-category taxonomy of explainable IDS
Fasla & Mehmood (2026) [34]	2016–2026	81	△	△	×	×	△	△	0	△	×	△	△	PRISMA review of XAI for web application threat detection; excludes network-level-only IDS
Nandhini et al. (2026) [6]	n.s.	n.s.	×	×	×	△	✓	△	0	△	×	✓	△	Comparative review of XAI methods for IDS on practical criteria; outlines a SOC deployment roadmap
Souto et al. (2026) [5]	2016–2026	71	✓	✓	×	△	✓	△	0	×	×	✓	✓	SLR of explainable IDS for IoT; per-study quality rubric and explanation-quality requirements
This review	2022–2026	109	✓	✓	✓	✓	△	✓	3	✓	△	✓	✓	PRISMA review of XAI-guided rule extraction in NIDS; per-study scoring, coding of deployment readiness and transferability, and inferential analysis

- Symbols used: ✓ = explicitly and systematically analysed; △ = discussed conceptually or partially; × = not addressed or not stated.
- Transferability dimensions: dataset, network environment, and attack type.
- [a]No search period reported, [b]Includes 14 survey papers.

[30, 31] or the lack of real-world deployment [34] as open problems. However, among other compared review papers, it is the only one whose rubric scores rule-extraction quality as a separate dimension and uses the dimension scores in inferential analysis. Its distinctive contribution lies in providing study-level descriptive and inferential evidence on these issues for XAI-guided rule extraction in network intrusion detection published between January 2022 and July 2026.

3 Methodology

3.1 Systematic Review Protocol

This systematic literature review based on guidelines from established PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework[35] to ensure rigorous, replicable, and transparent methodology. The review protocol was developed specifically to address the research objectives through systematic identification, selection, assessment, and synthesis of relevant literature at the intersection of explainable AI and deployable security rule extraction in network intrusion detection systems.

3.1.1 Search Criteria

The search strategy uses a combination of keyword searches across four academic databases: Scopus, IEEE Xplore, ACM Digital Library, and Web of Science. These databases were selected for their comprehensive coverage of computer science, network security, and artificial intelligence research. List of search queries for each database are summarized in Table 2. The search queries were designed to capture the intersection of three core concepts: explainable AI, network intrusion detection, and rule extraction/generation.

The search strings use Boolean operators to combine three concept clusters. The first cluster captures explainability: (“explainable AI” OR XAI OR “interpretable machine learning” OR interpretability). The second cluster targets rule generation and extraction: (“rule extraction” OR “rule generation” OR “rule mining” OR “feature extraction”). The third cluster covers intrusion detection and network security: (“intrusion detection” OR “network security” OR cybersecurity OR IDS OR NIDS).

Table 2 Database Search Queries

Database	Search Query
Scopus	TITLE-ABS-KEY((("explainable AI" OR xai OR "interpretable machine learning" OR "interpretability") AND ("rule extraction" OR "rule generation" OR "rule mining" OR "feature extraction")) AND ("intrusion detection" OR "network security" OR "cybersecurity" OR ids OR nids)) AND (LIMIT-TO(DOCTYPE,"ar") OR LIMIT-TO(DOCTYPE,"cp")) AND (LIMIT-TO(LANGUAGE,"English")) AND (LIMIT-TO(SUBJAREA,"COMP") OR LIMIT-TO(SUBJAREA,"ENGI"))
IEEE Xplore	("explainable AI" OR XAI OR "interpretable machine learning" OR interpretability) AND ("rule extraction" OR "rule generation" OR "rule mining" OR "feature extraction") AND ("intrusion detection" OR "network security" OR cybersecurity OR IDS OR NIDS)
Web of Science	TS=(("explainable AI" OR XAI OR "interpretable machine learning" OR "interpretability") AND ("rule extraction" OR "rule generation" OR "rule mining" OR "feature extraction") AND ("intrusion detection" OR "network security" OR "cybersecurity" OR IDS OR NIDS))
ACM Digital Library	Abstract:("explainable AI" OR XAI OR "interpretable machine learning" OR interpretability) AND Abstract:("rule extraction" OR "rule generation" OR "rule mining" OR "feature extraction") AND Abstract:("intrusion detection" OR "network security" OR cybersecurity OR IDS OR NIDS)

3.1.2 Inclusion and Exclusion Criteria

Based on each paper title and abstract, papers were included if they: (1) explicitly addressed explainability or interpretability of machine learning models in network intrusion detection contexts, (2) discussed rule extraction, rule generation, or the translation of model decisions into actionable security rules, (3) were published between January 2022 and July 2026, (4) were peer-reviewed conference papers or journal articles, and (5) achieved a Total_Score of at least 13 out of 20 in the quality assessment described in Section 3.1.3. A paper was excluded if any one of the conditions summarized in Table 3 applied.

3.1.3 Quality Assessment and Scoring System

Selected papers from selection of inclusion and exclusion stage were evaluated through four-category scoring system. In each category, a paper received a score ranged between 1 to 5 (Poor = 1, Limited = 2, Moderate = 3, Good = 4, Excellent = 5), they are:

- **XAI Technique Quality** (XAI_Score): usage of XAI in the research paper
- **Rule Extraction Quality** (Rule_Score): have explicit rules with clear IF-THEN formats or not
- **Network IDS Relevance** (Network_Score): use of network component to real traffic networks.
- **Methodological Rigor** (Method_Score): weak experimental design to rigorous protocols and reproducible.
- **Total Score** (Total_Score): sum of the four category, with a maximum scores 20 points.

The detailed scoring criteria are given in Table 4. Papers with a Total_Score below 13 were excluded from the final analysis to focus on research that makes substantive contributions; the effect of this threshold is examined in a sensitivity analysis (Section 3.3.3).

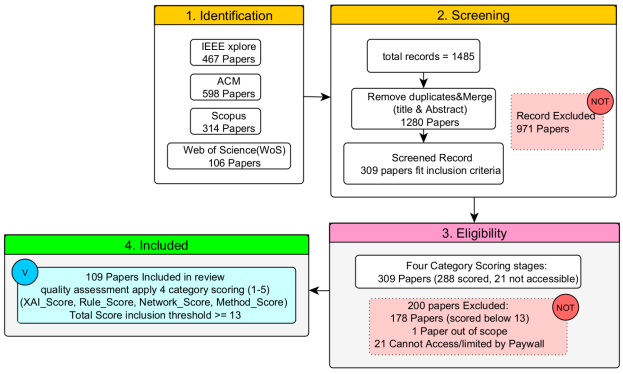


Fig. 1 PRISMA Flow Diagram for Systematic Literature Review

One included study [36] was authored by the first two authors of this review. It was screened against the same eligibility criteria and scored with the same rubric as all other studies. Excluding it does not change the main statistical findings.

Each scoring category were addressed for: XAI_Score addresses Objective 1, Rule_Score addresses Objective 2, Network_Score addresses Objective 3, and Method_Score addresses Objective 4. For Objective 2, deployment-ready and non-deployable papers are further compared using pearson point-biserial correlation and Cohen's d for the score components, together with a descriptive analysis of reported detection accuracy. For Objective 4, the Method_Score assessment is complemented by cross-category analysis and thematic analysis of the limitations reported in the reviewed papers.

3.1.4 Data Extraction Process

The entire PRISMA flow is shown in Figure 1. The searches returned 1485 records across the four databases. After duplicates were removed, 1280 unique records were screened by

Table 3 Exclusion Criteria (Any one excludes)

Criteria	Search exclusion
Out of Scope	– Pure host-based IDS focusing only on host/endpoint detection without network component – Non-security applications of ML interpretability in other domains (medical, finance, etc.) – Attack generation studies focusing on generating attacks rather than detecting them – Pure theoretical work with only mathematical proofs without implementation or evaluation
Methodological Issues	– No ML component: Traditional signature-based or rule-based systems without machine learning – No interpretability: Black-box ML approaches without any explainability component – No evaluation: Purely conceptual work without experimental validation – Survey/review only: Literature reviews without new technical contribution
Publication Type	– Short papers with less than 4 pages (insufficient detail for analysis) – Poster abstracts: Conference abstracts without full paper – Editorial/opinion pieces: Non-research content – Duplicate studies: Same study published in multiple venues
Quality Issues	– Non-peer reviewed: Workshop papers, technical reports, preprints without peer review – Language: Non-English publications – Accessibility: Cannot access full text after reasonable effort

Table 4 Four-Category Quality Assessment Scoring System

Category Score	XAI Technique Quality (XAI_Score)	Rule Extraction Quality (Rule_Score)	Network IDS Relevance (Network_Score)	Methodological Rigor (Method_Score)
5 (Excellent)	– multiple XAI techniques (SHAP + LIME + attention mechanisms) – both local and global explanations – novel XAI methods or creative applications – thorough validation and comparison	– explicit rule extraction algorithms – clear IF-THEN or decision tree formats – automated extraction processes – provided rule performance metrics	– real network traffic datasets, coverage of multiple attack types (DDoS + others) – practical IDS integration discussion – use of network-specific features	– rigorous experimental design – comparison with multiple baselines – statistical significance testing – reproducible implementation with code/data availability
4 (Good)	– established XAI techniques (SHAP or LIME) with clear methodology – proper implementation details, and some validation	– clear rule generation processes with structured formats (trees/patterns) – semi-automated extraction, and some validation	– standard network datasets (NSL-KDD, UNSW-NB15, CIC-IDS) – clear network attack detection focus – some IDS integration discussion, and analysis of network features	– clear experimental setup with baseline comparisons – comprehensive evaluation metrics (accuracy, precision, recall, F1) – acceptable implementation details
3 (Moderate)	– basic XAI techniques with some methodology and minimal validation	– described rule extraction with reasonable formats but manual or unclear processes	– basic network datasets with intrusion detection application but limited IDS integration	– basic experimental design with standard metrics and some implementation details
2 (Limited)	– vague XAI methods with poor implementation details and no validation	– implicit rule generation with unclear formats and vague processes	– mixed network/host detection with unclear network specificity	– weak experimental setup with no baselines and limited evaluation
1 (Poor)	– minimal XAI component with no clear methodology, merely mentioning "interpretability" without substantive explanation	– no explicit rules, only feature importance rankings, with no extraction methodology or evaluation	– no clear network component or IDS-specific application	– experimental design with no comparisons or implementation details

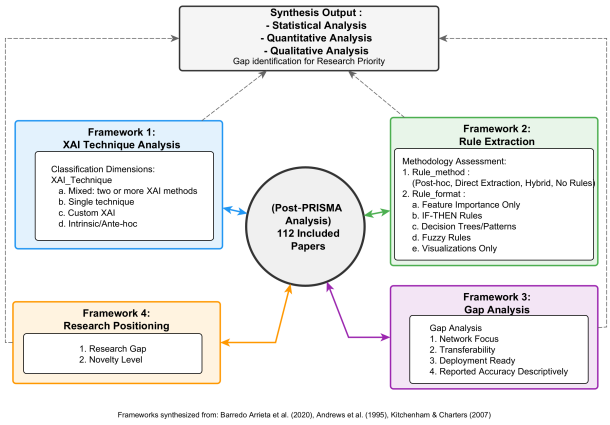


Fig. 2 Post PRISMA Analysis Methodology synthesized from [37][38][39]

title and abstract, and 971 records were excluded by the inclusion and exclusion criteria. Of the remaining 309 records, 21 full texts could not be accessed, and the other 288 full texts were assessed for eligibility and scored with the four-category system. Of these, 179 papers were excluded: 178 because their Total_Score was below 13 and 1 paper was out of scope (pure host-based detection without a network component, Table 3). The final included 109 papers for full extraction.

3.2 Analysis and Synthesis Methods

Table 5 summarizes the post-PRISMA guidelines used to extract detailed data from each selected papers. The guidelines were synthesized from established research [37–39] to cover the dimensions relevant to XAI implementation, rule extraction, gap analysis (limitations, transferability, and deployment readiness), and research positioning, as illustrated in Figure 2.

Data were analyzed using three approaches: (1) **Quantitative analysis** calculated frequency distributions of XAI techniques, rule extraction methods, network security application domains, and transferability levels, and summarized reported detection accuracy descriptively. (2) **Statistical analysis** used Pearson (point-biserial) correlation and Cohen's d to quantify the association between the quality scoring dimensions and deployment readiness, addressing Objective 2 (Rule Extraction) and Objective 4 (Methodological Gaps). (3) **Qualitative analysis** applied to the limitations reported in each study (the Research_Gap dimension in Table 5) and grouped related codes into research gap themes (Table 16), addressing Objective 4.

3.3 Statistical Analysis Methods

The statistical analysis examines which quality dimensions are associated with deployment readiness among the 109 included papers. All analyses used two-tailed tests with a conventional significance threshold of $\alpha = 0.05$ and were conducted in Python 3.12 with SciPy 1.11 and Pandas 2.1. All p-values are reported to four decimal places, or as $p < 0.0001$ for values below this threshold.

3.3.1 Variables and Operationalization

The analysis examines the relationship between two categories of variables:

Independent Variables (Scoring Dimensions). Each study was scored on four ordinal dimensions (1–5 scale) based on the quality assessment criteria in Table 4:

- *XAI_Score*: depth and quality of XAI technique implementation, from minimal mention of interpretability (1) to multiple validated XAI techniques with theoretical grounding (5).
- *Rule_Score*: rule extraction capability, from no explicit rules produced (1) to automated extraction with deployable IF-THEN formats and performance metrics (5).
- *Network_Score*: network security relevance, from no clear network component (1) to real network traffic with multiple attack types and IDS integration (5).
- *Method_Score*: methodological rigor, from minimal experimental design (1) to comprehensive reproducible protocols with statistical testing and baseline comparisons (5).
- *Total_Score*: sum of all four component scores (range 4–20, inclusion threshold ≥ 13).

Dependent Variable. *Deployment_Ready* is a binary variable from the Deployment_Ready classification in Table 5: Papers that classified as Fully Operational are coded 1, and papers that classified as Partial Solution or Theoretical Only are coded 0. Papers whose deployed artefact is a model or dashboard that outputs only feature importance, rather than rules, are coded 0. In total, $n_1 = 15$ studies are deployment-ready and $n_0 = 94$ are not.

Reported detection accuracy, extracted as defined in Table 5, is summarized descriptively and is not entered into the correlation analysis, because the values come from different datasets, data splits, and evaluation protocols and are therefore not directly comparable.

3.3.2 Statistical Methods and Justification

1. **Pearson's correlation coefficient** become measurement because of two reasons: (1) it is widely used in systematic review literature to interpreted correlation, as comparison with related papers. (2) the sample size ($N = 109$) is sufficiently large for the Central Limit Theorem to support

Table 5 Systematic Analysis: Post-PRISMA Data Extraction Guidelines

Analysis Dimension	Classification Categories and Descriptions
A. XAI IMPLEMENTATION CHARACTERISTICS	
XAI_Technique	**Mixed:** Implementations combining two or more XAI methods (such as SHAP + LIME, or SHAP + attention mechanisms).
	Single Technique: Approaches employing one established XAI method, including SHAP, LIME, attention weights, or gradient-based attribution (for example , Grad-CAM).
	Custom XAI: Explanation mechanisms or interpretable architectures designed by the authors for network intrusion detection.
	Intrinsic/Ante-hoc Interpretability: Interpretability inherent to a conventional model (for example, CART rules, Random Forest feature importance, or SVM decision boundaries) without any post-hoc XAI method or new explanation mechanism.
B. RULE EXTRACTION METHODOLOGY	
Rule_Method	**Post-hoc:** Explanation techniques applied to pre-trained black-box models to extract rules after model training.
	Direct Extraction: Rules derived directly from inherently interpretable architectures (decision trees, rule-based systems) where model structure naturally provides rules.
	Hybrid: Methodologies that combine an interpretable model (such as a decision tree or rule-based system) with post-hoc XAI techniques (such as SHAP, LIME, or Anchors) to balance interpretability and performance.
	No Rules: Papers implementing XAI techniques but producing no explicit extractable rules whatsoever.
Rule_Format	**Feature Importance Only:** Outputs limited to ranked lists of important features without actionable rule structures.
	IF-THEN Rules: Structured conditional rules in standard IF-THEN format directly deployable in rule engines and firewalls.
	Decision Trees/Patterns: Tree-based structures or attack patterns convertible to operational filtering formats.
	Fuzzy Rules: Fuzzy logic-based rules with linguistic variables and membership functions.
	Visualizations Only: Explanatory outputs remaining as visual representations without systematic translation to deployable security rules.
C. GAP ANALYSIS: DEPLOYMENT VIABILITY ASSESSMENT	
Network_Focus	**General IDS:** Addressing intrusion detection covering multiple attack types (DDoS, PortScan, Brute Force, Web attacks, etc.) without specific focus.
	DDoS-Specific: Research focusing specifically on Distributed Denial of Service attacks as the primary research problem.
	IoT/Industry-Specific: Specialized approaches for specific network environments (IoT, Industrial Control Systems, etc.).
	Diverse Scenarios: Approaches applied across multiple distinct network intrusion detection scenarios with cross-domain validation.
	Non-IDS (Malware/Botnet): Studies whose primary task is malware or botnet detection or classification rather than general network intrusion detection.
Reported_Accuracy	Highest overall detection accuracy reported for the proposed model on any dataset or, when no single model is proposed, the highest among the models evaluated by the authors. Accuracy values of baselines cited from other works are not used. Studies that report only other metrics (for example , F1-score, false-positive rate, or rule fidelity) are coded as *not reported*.
Transferability	**Cross-Dataset:** Testing on multiple benchmark datasets.
	Cross-Network Environment: Validation across different network contexts (enterprise, cloud, IoT, industrial systems).
	Cross-Attack Type: Effectiveness demonstrated against diverse attack categories (DDoS, reconnaissance, malware, etc.).
	Each study is coded for the presence of each dimension, and the number of dimensions present defines its transferability level: No Validation (0), 1-D, 2-D, or 3-D. A 3-D study is validated along all three dimensions, not merely on three datasets.
Deployment_Ready	**Fully Operational:** Papers providing complete operational rules in deployable formats (IF-THEN, decision trees, patterns) with actual deployment validation or clear integration pathways.
	Partial Solution: Offering XAI explanations, feature rankings, or pattern descriptions without systematic rule translation or deployment validation.
	Theoretical Only: Contributions lacking any practical path to operational implementation, remaining at conceptual or experimental stage.
	For statistical analysis, Fully Operational is coded 1 and Partial Solution or Theoretical Only is coded 0. Studies whose deployed artefact is a model or dashboard that outputs only feature importance, rather than rules, are classified as Partial Solution.
D. RESEARCH POSITIONING	
Research_Gap	**Explicitly Stated Limitations:** Gaps acknowledged by authors (computational cost, single dataset limitation, etc.).
	Methodological Blind Spots: Identified limitations not addressed by authors, particularly: - Adaptive rule update mechanisms - Cross-domain transferability validation - Rule maintenance and evolution strategies - Performance degradation over time
	Deployment Barriers: Practical obstacles preventing real-world implementation (computational overhead, integration complexity, scalability issues).

parametric methods even with ordinal input data. Because the dependent variable (Deployment_Ready) coded as binary (0 or 1), the Pearson correlation reduces to the point-biserial correlation (r_{pb}), a method for continuous–binary relationships.

To assess the association between each scoring dimension (x_i) and deployment status (y_i), we computed:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

where x_i represents a paper's score on a given dimension (for example , Rule_Score), y_i represents its deployment status (1 or 0), $\bar{x}$ and $\bar{y}$ are sample means, and $n = 109$ is

Table 6 Interpretation Guidelines for Effect Sizes

Measure	Small	Medium	Large
Cohen's d	0.20–0.49	0.50–0.79	≥ 0.80
Pearson's r	0.10–0.29	0.30–0.49	≥ 0.50

the total number of papers. For each correlation, statistical significance was evaluated using a t-test with $n-2$ degrees of freedom:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{2}$$

The coefficient of determination ($R^2 = r^2$) quantifies the proportion of deployment variance explained by each scoring dimension.

2. **Cohen's d (Effect Size for Group Differences)** Cohen's d [40] was used to quantify whether deployment-ready papers score meaningfully higher on Rule_Score than non-deployable papers, expressed as the group difference in standard deviation units:

$$d = \frac{M_{deployed} - M_{non\text{-}deployed}}{s_{pooled}} \tag{3}$$

where $M_{deployed}$ is the mean Rule_Score of deployment-ready papers ($n_1 = 15$), $M_{non\text{-}deployed}$ is the mean Rule_Score of non-deployable studies ($n_0 = 94$), and s_{pooled} is the pooled standard deviation:

$$s_{pooled} = \sqrt{\frac{s_1^2(n_1-1) + s_0^2(n_0-1)}{n_1+n_0-2}} \tag{4}$$

Table 6 provides interpretation guidelines following conventional thresholds [40], where $|d| < 0.2$ represents negligible effect, $0.2 \leq |d| < 0.5$ small effect, $0.5 \leq |d| < 0.8$ medium effect, and $|d| \geq 0.8$ large effect.

3.3.3 Sensitivity Analysis

Sensitivity Analysis conducted by repeating the correlation and effect-size analyses test under two alternatives:
(i) a lower quality threshold of Total_Score ≥ 12 for papers published between January 2022 and July 2026, which adds the 12 papers with a Total_Score of 12 for which deployment labels were available ($n = 121$), and (ii) the same threshold with the 16 studies published before January 2022 that had been scored with the same rubric added ($n = 137$). The results are reported in Section 4.3.1.

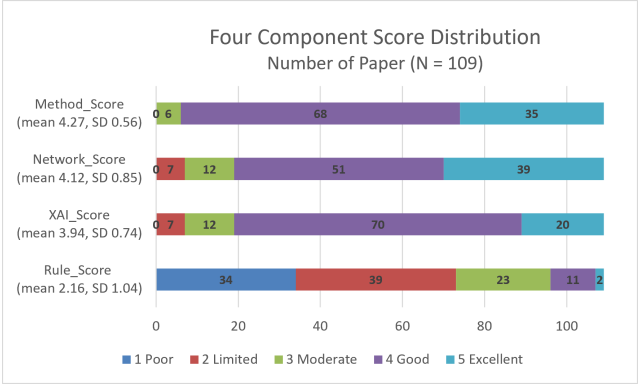


Fig. 3 Distribution of the four component scores across the 109 included papers

4 Results and Discussion

The 109 included papers have a mean Total_Score of 14.49/20 (SD = 1.44): 58 papers (53.2%) score 13–14, 49 (45.0%) score 15–17, and only 2 (1.8%) reach the High range of 18–20. Because every included paper meets the inclusion threshold of 13 points, this distribution describes the variation of capable methodologically among papers.

Figure 3 displays the unevenly quality spread across the four components. Most papers score 4 or 5 on methodological rigor (103 papers; mean 4.27/5), network security relevance (90; mean 4.12), and XAI implementation (90; mean 3.94), whereas 73 papers (67.0%) score 2 or lower on rule extraction (mean 2.16/5), which quantifies the rule translation gap. Rule_Score is also the most widely spread component (SD = 1.04), indicates that no established standard exists for translating XAI outputs into operational rules.

4.1 Hierarchical Taxonomy of XAI-NIDS Approaches

Based on Post-PRISMA Data extraction guidelines from Table 5, the systematic analysis classified into three level of XAI-NIDS approaches, as displayed in Table 7. The 109 included papers were classified along three dimensions: XAI technique category, rule generation capability, and transferability validation. There are five main findings across the reviewed papers:

1. **XAI Category Dominance: Mixed and Single-Technique Approaches.** SHAP-based methodologies dominates current intrusion detection research and is used in 77 papers (70.6%): 33 papers (30.3%) use SHAP as their only XAI technique, and 44 of the 47 mixed-method papers combine SHAP with other techniques. Frameworks such as XAI-IDS use SHAP feature rankings to identify the features that drive detection decisions [1]. Single techniques other than SHAP are rare: LIME-only approaches appear in 4 papers (3.7%), attention-based approaches in 3 pa-

Table 7 Hierarchical Taxonomy of XAI-NIDS Approaches

Level	Category	N	%	Description	Example Papers
LEVEL 1: XAI Technique Category					
1.1	Mixed Methods	47	43.1%	Two or more XAI techniques (44 of 47 include SHAP)	[1], [41], [42], [43], [44]
1.2	Single-Technique	41	37.6%	One established XAI method	
	SHAP-only	*33*		*Most common single technique*	[45], [46], [47], [48], [49]
	LIME-only	*4*		*Local surrogate explanations*	[50], [51]
	Attention-based	*3*		*Attention weights as explanation*	[52], [53]
	Gradient-based	*1*		*Attribution methods*	[54]
1.3	Custom Methods	20	18.3%	Explanation mechanisms or interpretable architectures designed by the authors	[2], [11], [55], [56], [57], [58]
1.4	Intrinsic/Ante-hoc	1	0.9%	Built-in interpretability of a conventional model	[59]
LEVEL 2: Rule Generation Capability					
2.1	Deployment-Ready	15	13.8%	Operational rules in deployable formats	
	Decision trees/patterns	*9*		*Tree paths or traffic patterns*	[58], [60], [56], [57], [52]
	IF-THEN rules	*6*		*Conditional rules for rule engines*	[11], [55], [61], [44], [49], [36]
2.2	Non-Deployable	94	86.2%		
	No extractable rules	*42*		*Feature importance or visualizations only*	[12], [62], [63], [64], [65]
	Partial deployment	*52*		*Rules or feature outputs without deployment validation or integration*	[1], [2], [43], [66], [67]
LEVEL 3: Transferability Validation (Cross-Dataset, Cross-Network Environment, Cross-Attack Type)					
3.1	Single-Dimension	55	50.5%	Dataset (33), network environment (9), or attack type (13) only	[2], [57], [55], [66]
3.2	Two-Dimension	16	14.7%	Mostly dataset + network (10 papers)	[46], [56], [43], [54], [68]
3.3	Three-Dimension	5	4.6%	Dataset + network + attack type	[69], [70], [71], [72], [73]
3.4	No Validation	33	30.3%	No transferability validation	[41], [74], [44], [49]

pers (2.8%), and gradient-based attribution only 1 paper (0.9%).

2. **Custom Method Limitations.** Custom XAI methods appear in 20 papers (18.3%), including set-theoretic interpretability mechanisms [2], abnormal logical representation learning [11], and tree distillation of an anomaly detector with gradient-based augmentation [58]. Custom methods have the highest deployment rate of the XAI categories (Table 11), however each research is designed for a specific setting, they lack standardization, which limits cross-study comparison and adoption in production environments.
3. **Attention Mechanism Underutilization.** Attention-based explanation is the primary XAI technique in only 3 papers (2.8%), such as the network transformer of Marino et al. [52] and the SRU–TabNet model of Chen et al. [53]. Only Marino et al. produces deployment-ready rules; in the other two papers, the attention weights remain explanations rather than inputs to rule extraction.
4. **Rule Generation Capability (Level 2).** Only 15 papers (13.8%) produce deployment-ready rules: 9 as decision trees or patterns and 6 as IF-THEN rules. Of the 94 non-deployable papers, 42 papers(38.5% of the corpus) produce no extractable rules, and 52 papers (47.7%) produce rules or feature outputs without deployment validation or an integration pathway.
5. **Transferability Validation (Level 3)** remains limited: 33 papers (30.3%) report no transferability validation, 55 papers(50.5%) validate along a single dimension, and only 5 papers (4.6%) validate across all three dimensions (dataset, network environment, and attack type).

4.2 Transferability Gap

As shown in Level 3 of Table 7 there are three transferability dimensions (cross-dataset, cross-network environment, and cross-attack type). 33 papers (30.3%) report no transferability validation and 55 papers (50.5%) validate along a single dimension. Papers that evaluate several benchmark datasets, such as NSL-KDD, UNSW-NB15, and CIC-IDS2017, mostly do not test the same approach across network environments or attack categories [1, 2, 57]. Even one of the few papers that trains on one dataset and evaluates on another reports that performance "slightly degrades" under this shift [68].

Two-dimension validation appears in 16 papers (14.7%), 10 papers across datasets and network environments, then 5 papers validated by datasets and attack types and only 1 paper validated by network environments. For examples the evaluation of two standard feature sets on three datasets built in different network environments [46], the transfer of GGFAST analyzers to encrypted SSH and TLS traffic [56], and the selective transfer of a trained encoder from wired and IoT traffic to a wireless dataset [54].

Only 5 papers (4.6%) use three-dimension validation, all published in 2026 [69–73], none of them is deployment-ready. A contrast between 55 single-dimension papers to 5 three-dimension papers indicates that researchers rarely validate their approaches beyond the conditions used for training. Without cross-environment validation, deployed rules remain vulnerable to changes in the operational environment.

Detailed of transferability validation papers breakdown presented in Table 8.

4.3 Correlation Analysis Results

Table 9 reports the correlation between each scoring variable (XAI_Score, Rule_Score, Network_Score, Method_Score, Total_Score, and the number of validated transferability dimensions) and deployment readiness. The correlations are computed from all 109 included papers, with each paper coded 1 if it is deployment-ready and 0 if not. The correlation analysis reveals three findings:

1. **The Accuracy–Deployment Disconnect.**
   High detection accuracy is nearly universal and is not related to deployment readiness (Table 10). Only 7 of the 58 papers with at least 99% accuracy (12.1%) are deployment-ready, and 23 of them (39.7%) produce no extractable rules. Because accuracy values are not comparable across datasets and evaluation protocols, they are summarized descriptively only. The deployment gap is therefore not a matter of better detection models, but a methodological problem of translating models and their explanations into rules.
2. **Rule_Score Is Most Strongly Associated with Deployment Readiness.**
   Rule extraction quality shows the strongest association with deployment readiness ($r = 0.455$, $p < 0.0001$), the corpus has variance of $R^2 = 20.7\%$ in deployment status, with a large effect of standardized mean difference (Cohen's $d = 1.47$): confirm that mean deployment-ready papers where 69% higher compared to non-deployable papers (mean Rule_Score 3.33 vs 1.97 on the 1–5 scale) . XAI_Score, Method_Score, Network_Score, and the number of transferability dimensions show no significant association ($p > 0.15$ for all four). Even if Extractable rules are required for deployment but do not guarantee it: none of the 42 papers without extractable rules is deployment-ready, yet 23 of the 36 papers with a Rule_Score of 3 or higher are still not deployment-ready.
3. **XAI Method Independence.**
   XAI technique choice shows a weak association with deployment readiness (Table 11). Custom methods have the highest deployment rate (7 of 20 papers, 35.0%), followed by Mixed (5 of 47 papers, 10.6%) and Single-technique approaches (3 of 41 papers, 7.3%); SHAP-only approaches reach 6.1% (2 of 33 papers). Papers that use SHAP only 9.1% that deployment-ready(7 of 77 papers). Deployment readiness therefore depends on rule extraction methodology rather than on the choice of XAI technique.

The three findings show that only Rule_Score is the one most strongly associated with deployment readiness, which identifies rule translation as critical success factor and the primary research gap.

4.3.1 Sensitivity Analysis Results

Table 12 repeats the correlation and effect-size analyses under the alternative corpus definitions described in Section 3.3.3. Rule_Score remains significantly associated with deployment readiness in every variant ($r = 0.455$ to 0.518, all $p < 0.0001$; $d = 1.47$ to 1.67), whereas XAI_Score remains non-significant ($r = 0.049$ to 0.138, all $p > 0.15$). The main findings therefore do not depend on the quality threshold or on the exclusion of papers published before 2022, and the deployment gap reflects a specific blind spot in rule extraction rather than insufficient research quality.

4.4 The Rule Extraction Challenge

Rule extraction method (Rule_Method in Table 5) were grouped by Table 13. It shows the gap between displaying explanations and producing deployable security rules, they are:

1. **No Rules.** Forty-two papers (38.5%) that using XAI not produce extractable rules. Most of them use established

Table 8 Transferability Validation Dimensions Across Papers

Paper Reference	Validation Description	Dataset	Network	Attack
Papers validated along all three dimensions (all 5 papers)				
Saiyed & Al-Anbagi (2026)[69]	HL-IoT, ToN-IoT, and Edge-IIoTSet; IoT and IIoT environments; high- and low-volume DDoS variants; datasets evaluated independently	✓	✓	✓
Wilson et al. (2026)[70]	CICIDS2018, CIC-ToN-IoT, and NF-UNSW-NB15-v2; enterprise, IoT, and hybrid settings; per-class results; datasets evaluated independently	✓	✓	✓
Ekpo et al. (2026)[71]	Five datasets (VANET, IoV, 5G, Edge-IIoT, cloud); diverse attack families; datasets evaluated independently	✓	✓	✓
Dresch et al. (2026)[72]	X-CANIDS, CICIDS, and CICIoMT; vehicular, conventional, and healthcare IoMT contexts; six attack types; methods re-run per dataset	✓	✓	✓
Onuorah et al. (2026)[73]	CSE-CIC-IDS2018, CIC-DDoS2019, and TII-SSRC-23; enterprise/cloud, DDoS, and Industry 4.0 IoT environments; datasets evaluated independently	✓	✓	✓
Papers validated along two dimensions (sample)				
Sarhan et al. (2022)[46]	Two feature sets evaluated on three datasets built in different network environments	✓	✓	–
Piet et al. (2023)[56]	Cross-dataset evaluation; analyzers transferred to encrypted SSH and TLS traffic	✓	✓	–
Zouhri & Idri (2025)[43]	Three datasets across multiple network environments; concept drift analysis	✓	✓	–
Kara & Boyaci (2026)[54]	Encoder trained on CICIDS2018 and CICIoT2023 transferred to AWID3 (wired-to-wireless and IoT-to-wireless)	✓	✓	–
Shafir et al. (2026)[68]	Trained on CICIDS2017 and tested on CICDDoS2019, and vice versa; more than 35 attacks	✓	–	✓
Uccello & Nadjm-Tehrani (2026)[75]	Three 5G network contexts; multiple attack categories; a separate model per dataset	–	✓	✓
Papers validated along a single dimension (sample)				
Pai et al. (2024)[2]	NSL-KDD, UNSW-NB15, and UKM-IDS20	✓	–	–
Fan et al. (2025)[57]	CIC IoT 2022, UNSW-NB15, IoT Network Intrusion, and power grid traffic	✓	–	–
Wichtlhuber et al. (2022)[55]	Different IXPs and geographic locations	–	✓	–
Ayoughi et al. (2024)[66]	Single industrial system; DoS versus DDoS	–	–	✓
Kusuma & Idris (2026)[36]	Five DDoS categories (NetBIOS, UDP, MSSQL, LDAP, SYN) within one dataset	–	–	✓
Papers without transferability validation (sample)				
Garroppo et al. (2025)[74]	Single dataset	–	–	–
Saied & Guirguis (2025)[44]	Single N-BaIoT dataset	–	–	–
Namjoo et al. (2025)[49]	Single dataset	–	–	–

Table 9 Correlation Analysis: Scoring Dimensions vs Deployment Readiness

Variable	r	p	R^2	Interpretation
Rule_Score	0.455	<0.0001	20.7%	Medium; significant
Total_Score[a]	0.384	<0.0001	14.7%	Medium; significant
XAI_Score	0.138	0.1533	1.9%	Not significant
Method_Score	0.097	0.3165	0.9%	Not significant
Network_Score	−0.088	0.3622	0.8%	Not significant
Transferability[b]	−0.035	0.7193	0.1%	Not significant
Cohen's d Rule_Score	Deployment-ready: (M=3.33, n=15) vs Non-deployable: (M=1.97, n=94)			d = 1.47; large effect; 69% higher

techniques (20 mixed-method and 17 single-technique papers) and stop at feature-level explanations [12, 62, 63]. None of them is deployment-ready.

2. **Post-hoc.** Post-hoc explanation of trained black-box models is the most common method (45 papers, 41.3%). Yet 35 of the 45 output only feature importance, leaving security administrators with a list of suspicious features but no logic to implement [1, 43]. The 7 deployment-ready post-hoc papers convert explanations into trees or patterns, such as the global explanations of GEAD [58] and the allow-list rules extracted from an autoencoder [57].
3. **Direct extraction.** Direct extraction reads rules from interpretable models such as decision trees or rule-based systems. It is used in 11 papers (10.1%) and has the highest deployment rate (54.5%), and the papers that report accuracy reach a median of 99.57%, which shows that interpretability need not cost detection performance. Ex-

Table 10 Reported Detection Accuracy by Deployment Status and Rule Extraction Method

Group	N	Reported n (%)	$\geq$ 99% n (% of reported)	Median (%)	Not reported n (%)
All papers	109	89 (81.7%)	58 (65.2%)	99.47	20 (18.3%)
By deployment status					
Deployment-ready	15	10 (66.7%)	7 (70.0%)	99.12	5 (33.3%)
Non-deployable	94	79 (84.0%)	51 (64.6%)	99.47	15 (16.0%)
By rule extraction method					
No Rules	42	35 (83.3%)	23 (65.7%)	99.47	7 (16.7%)
Post-hoc	45	37 (82.2%)	25 (67.6%)	99.56	8 (17.8%)
Direct Extraction	11	8 (72.7%)	5 (62.5%)	99.57	3 (27.3%)
Hybrid	11	9 (81.8%)	5 (55.6%)	99.00	2 (18.2%)

Accuracy is the highest overall detection accuracy reported for the proposed model (Table 5)

Table 11 Deployment Readiness by XAI Method Category

XAI Method	Deploy	Non-Deploy	Rate
Custom	7	13	35.0%
Mixed	5	42	10.6%
Single	3	38	7.3%
of which SHAP-only	2	31	6.1%
Intrinsic/Ante-hoc	0	1	0.0%
Total	**15**	**94**	**13.8%**

Table 12 Sensitivity Analysis: Association of Rule_Score and XAI_Score with Deployment Readiness

Corpus definition	n (n_1)	Rule_Score r	Cohen's d	XAI_Score r (p)
Main: Total_Score $\geq$ 13 (Jan 2022–Jul 2026)	109 (15)	0.455	1.47	0.138 (0.1533)
Total_Score $\geq$ 12 (Jan 2022–Jul 2026)	121 (16)	0.489	1.64	0.107 (0.2440)
Total_Score $\geq$ 12 (incl. papers before 2022)	137 (21)	0.518	1.67	0.049 (0.5711)

All Rule_Score correlations: $p < 0.0001$. XAI_Score: r with its p-value in parentheses. n_1 = number of deployment-ready papers.

amples include IF-THEN rules generated from a neural-symbolic architecture [11], set-theoretic operations [2], and firewall rules derived from Weight of Evidence encoding [55].

4. **Hybrid.** Hybrid methods combine an interpretable model, such as a decision tree, with post-hoc XAI techniques, such as SHAP. They appear in 11 papers (10.1%) and reach a high mean Rule_Score (3.18), but only 2 of them are deployment-ready: SHAP and Anchors for rule mining in a SIEM [61] and the translation of SHAP attributions into Suricata signatures [36]. Other hybrid designs, such as ML-enhanced automata learning [66] and explainable-by-design ensembles [67], produce rules or patterns without deployment validation.

4.5 The Deployment and Operational Limitations

Table 14 shows the details of the 15 deployment-ready papers by rule format, XAI method, extraction method, network

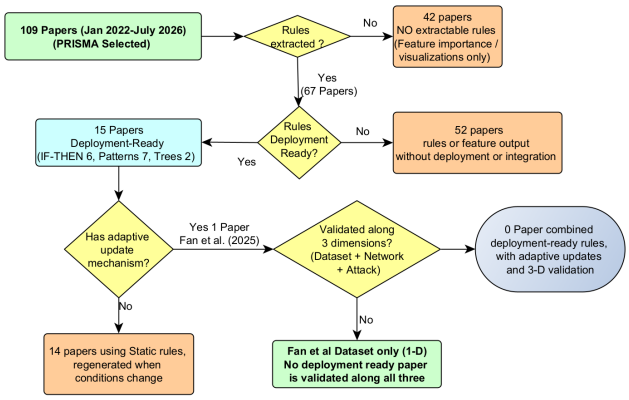


Fig. 4 Deployment readiness filter. "Rules extracted" follows the Rule_Method coding in Table 5.

focus, and transferability. Six papers produce IF-THEN rules and nine produce decision-tree or pattern-based rules. Only one deployment-ready paper, Fan et al. [57], provides an update strategy for its rule library, using custom gradient ascent exploration as its XAI method. The other 14 papers would require their rules to be regenerated when network conditions change, which makes them hard to maintain in the long run. Fan et al. is also the only paper among the 109 reviewed papers that includes an adaptive update mechanism (1 of 109, 0.9%).

The deployment-ready solutions are also limited in transferability: five papers report no transferability validation, seven validate along a single dimension (1-D), and only three validate along two dimensions (2-D) [11, 56, 58]. Figure 4 shows how the number of papers narrows at each requirement: no paper combines deployment-ready rules, adaptive updates, and validation along all three dimensions.

4.6 Network Focus Research

From the network focus point of view, 16 papers (14.7%) focus on Distributed Denial of Service (DDoS) attacks as their primary objective (Table 15). Most of them apply established XAI techniques: 6 use a single technique, 7 use mixed methods, 2 use custom methods, and 1 relies on intrinsic interpretability. Only 2 of the 16 are deployment-ready: IXP Scrubber, which produces filtering rules for DDoS traffic at Internet Exchange Points [55], and the translation of SHAP attributions into Suricata signatures [36]. Validation is also narrow: 6 papers report no transferability validation, 8 validate along a single dimension, and only 2 validate along more than one dimension, one along two and one along all three [69]; neither of these two produces extractable rules.

Among papers outside the DDoS-specific category, the interpretable generalization mechanism of Pai et al. [2] reaches a Rule_Score of 4 (on the 1–5 scale) with a clear pattern extraction methodology and is validated on three datasets

Table 13 Methodological Categories in XAI-NIDS Research

Category	Papers	Deploy Rate	Mean Rule_Score	Characteristic Pattern	Representative Examples
No Rules	42 (38.5%)	0.0%	1.55	XAI outputs end at feature importance or visualizations; no rule conversion methodology	X-GANet[12], Hybrid DL with SHAP/LIME[62], XAI consensus[63]
Post-hoc	45 (41.3%)	15.6%	2.22	XAI applied to trained black-box models; 35 of 45 output feature importance only	XAI-IDS[1], CTGAN-ENN[43], GEAD[58]
Direct Extraction	11 (10.1%)	54.5%	3.18	Rules read directly from interpretable models (e.g., decision trees); often custom designs	ALRL[11], IXP Scrubber[55], GlassBoost[49]
Hybrid	11 (10.1%)	18.2%	3.18	Interpretable models combined with post-hoc XAI techniques (e.g., SHAP); deployment still rare	SHAP+Anchors in SIEM[61], Automata learning[66], XAI-to-Suricata[36]

Table 14 Deployment-Ready Papers by Rule Format and Extraction Methodology

Authors	Rule Format	XAI Method	Extraction Method	Network Focus	Transfer
IF-THEN Rules (6)					
Wichtlhuber et al. (2022)[55]	IF-THEN	Custom (WoE)	Direct	DDoS	1-D
Xu et al. (2024)[11]	IF-THEN	Custom (logic)	Direct	General IDS	2-D
Uccello et al. (2024)[61]	IF-THEN	Mixed (SHAP+Anchors)	Hybrid	General IDS	1-D
Saied & Guirguis (2025)[44]	IF-THEN	Mixed (SHAP+LIME+RF rules)	Direct	IoT/Industry	None
Namjoo et al. (2025)[49]	IF-THEN	SHAP-only (TreeSHAP)	Direct	General IDS	None
Kusuma & Idris (2026)[36]	IF-THEN	Mixed (TreeSHAP+DeepSHAP)	Hybrid	DDoS	1-D
Decision Trees/Patterns (9)					
Piet et al. (2023)[56]	Patterns	Custom (snippets)	Direct	General IDS	2-D
Han et al. (2024)[58]	Trees	Custom (tree distillation)	Post-hoc	General IDS	2-D
Li et al. (2024)[60]	Trees	Custom (tree regularization)	Post-hoc	General IDS	1-D
Baral et al. (2024)[42]	Patterns	Mixed (SHAP+LIME)	Post-hoc	IoT/Industry	None
Lee et al. (2024)[48]	Patterns	SHAP-only	Post-hoc	General IDS	1-D
Garroppo et al. (2025)[74]	Patterns	Custom (Grad-CAM)	Post-hoc	IoT/Industry	None
Marino et al. (2025)[52]	Patterns	Attention	Direct	General IDS	1-D
Fan et al. (2025)*[57]	Patterns	Custom (gradient ascent)	Post-hoc	IoT/Industry	1-D
Paltun et al. (2025)[76]	Patterns	Mixed (SHAP+LIME+PI)	Post-hoc	General IDS	None

*Only paper in the selected corpus with an adaptive update mechanism. Transfer = number of transferability dimensions validated (None = no validation). PI = permutation importance; RF = random forest; WoE = Weight of Evidence.

(NSL-KDD, UNSW-NB15, and UKM-IDS20), but it is not deployment-ready and provides no adaptation mechanism for evolving threats. Similarly, the GGFAST framework [56] automatically generates interpretable network classifiers and is validated across datasets and network environments, yet it also lacks adaptive capabilities. This pattern repeats across several papers: they achieve excellent results in isolated settings but do not address the complete operational requirements of deployable and adaptive security systems.

The detailed network focus distribution is shown in Table 15.

Table 15 Network Focus Distribution

Network Focus Category	Count	Percentage	Deployment ready
General IDS	46	42.2%	9 (19.6%)
IoT/Industry-Specific	33	30.3%	4 (12.1%)
DDoS-Specific	16	14.7%	2 (12.5%)
Diverse Scenarios	8	7.3%	0 (0.0%)
Non-IDS (Malware/Botnet)	6	5.5%	0 (0.0%)
Total	**109**	**100.0%**	**15 (13.8%)**

4.7 Research Gap Patterns

Each paper's Research_Gap entry was restricted to limitations stated by the authors and related codes were grouped into themes (Table 16). Because authors do not always state

Table 16 Research Gap Themes Across the Included Papers

Research Gap Theme	Codes	Papers	% of 109
XAI-to-rule translation gap	3	59	54.1%
Rule adaptation and robustness gap	2	22	20.2%
Methodological issues (class imbalance, expert dependence, evaluation standardization, privacy)	4	21	19.3%
Cross-domain validation absence	1	20	18.3%
Computational overhead and scalability	1	16	14.7%
Papers with at least one coded limitation	11	95	87.2%
A paper can contribute to more than one theme			

the limitations of their own work, each theme was also compared with the extraction fields (rule extraction method, rule format, deployment readiness, and transferability), so that stated and observed limitations can be compared directly. Of the 109 papers, 95 received at least one code, and a paper can contribute to more than one theme. The themes are:

1. **Recurring themes and under-reporting.** The XAI-to-rule translation gap is the most frequently acknowledged theme (Table 16): authors recognize that feature-importance outputs are not converted into operational rules or validated in a running IDS. The other themes cover concept drift and untested adversarial robustness [77–79], narrower methodological issues such as class imbalance and expert dependence, performance degradation across datasets [68], and the cost of computing SHAP explanations [80, 81]. The extraction fields show that the translation gap is nevertheless under-reported: only 44 of the 94 non-deployable papers (46.8%) acknowledge it, and only 18 of the 42 papers that produce no rules at all (42.9%). The field's largest structural deficit is therefore one that most of the affected papers do not name as a limitation.
2. **Trajectory over time.** Figure 5 groups the papers by publication period. SHAP use rises from 60.9% in 2022–2023 to 77.8% in 2026, and the share of papers without extractable rules rises from 13.0% to 70.4%, whereas deployment readiness stays low and falls to 3.7% in 2026 (one paper [36]). Mean Rule_Score falls from a peak of 2.44 in 2024 to 1.67 in 2026, although publication year is not significantly correlated with Rule_Score (Pearson's $r = -0.111$, $p = 0.2492$) or with deployment readiness (Pearson's $r = -0.045$, $p = 0.6418$). Only validation breadth improves: 5 of the 27 papers published in 2026 (18.5%) validate along all three transferability dimensions.
3. **Closest cases.** Table 17 lists of the 5 papers that come closest to XAI-guided, deployable, transferable, and adaptive rule extraction. Each provides some of the required capabilities but lacks at least one, so every capability exists somewhere in the corpus but never in a single system.

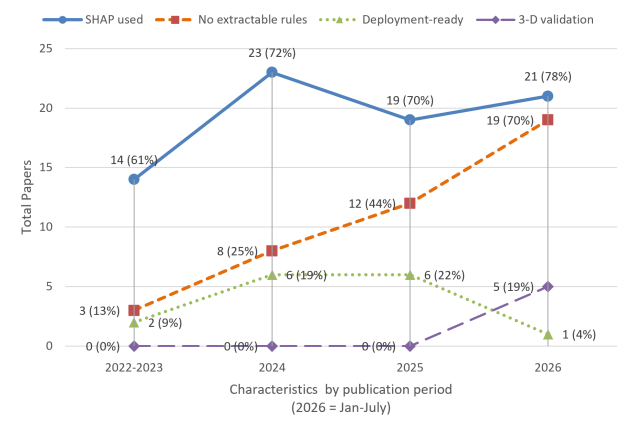


Fig. 5 Characteristics of included papers by publication period.

Table 17 Papers Closest to XAI-Guided, Deployable, Transferable, and Adaptive Rule Extraction

Paper	Rules produced	Dep.	Trans.	Adapt.	Missing link
Fan et al. (2025)[57]	Allow-list rule library	✓	1-D	✓	Validated across datasets only
Han et al. (2024)[58]	Regression-tree rules distilled from an anomaly detector	✓	2-D	×	Rules are re-extracted rather than updated; no adaptive mechanism
Ding et al. (2026)[82]	79 statistical IF-THEN rules and 236 RF path rules	×	1-D	×	Rules not deployed; single dataset
Kusuma & Idris (2026)[36]	12 Suricata signatures	✓	1-D	×	79 benign test flows, too few to validate false positives; single dataset
Qian & He (2026)[83]	IF-THEN belief rule base	×	None	×	Expert-assigned priors; single dataset; no IDS integration

Dep. = deployment-ready; Trans. = number of transferability dimensions validated; Adapt. = adaptive update mechanism; RF = random forest.

5 Conclusion

5.1 Summary of Key Findings

Through the analysis of 109 papers published between January 2022 and July 2026, this systematic review answers the four objectives:

1. **Objective 1 (Systematic Analysis of XAI Techniques).** The hierarchical taxonomy shows that SHAP dominates current research and is used in 77 papers (70.6%), either alone (33 SHAP-only papers, 30.3%) or combined with other techniques. Mixed methods account for 47 papers (43.1%), single-technique approaches for 41 (37.6%), custom methods for 20 (18.3%), and intrinsic interpretability for 1 (0.9%). Attention-based explanation is the primary technique in only 3 papers (2.8%). XAI implementation quality is high (mean XAI_Score 3.94/5) but is not significantly associated with deployment readiness (Pearson's correlation coefficient $r = 0.138$, p-value $= 0.1533$). Custom methods reach the highest deployment rate (35.0%), whereas SHAP-only approaches reach

only 6.1%, which shows that neither the choice nor the number of XAI techniques determines deployment outcomes by itself.

2. **Objective 2 (Evaluation of Rule Extraction Methods).** The analysis reveals a clear gap between explanation and operational deployment:
   - **No Extractable Rules:** 42 papers (38.5%) produce no extractable rules despite implementing XAI techniques, and 79 papers (72.5%) output only feature importance.
   - **Deployment Gap:** Only 15 papers (13.8%) produce deployment-ready rules, as IF-THEN rules (6 papers) or decision trees and patterns (9 papers). The remaining 94 papers (86.2%) are not deployment-ready: 42 produce no rules, and 52 (47.7%) produce rules or feature outputs without deployment validation or an integration pathway.
   - **Accuracy–Deployment Disconnect:** High detection accuracy does not translate into deployable rules. Of the 89 papers that report accuracy, 58 (65.2%) reach at least 99%, yet only 7 of these are deployment-ready. Rule extraction quality, by contrast, is the dimension most strongly associated with deployment readiness ($r = 0.455$, $p < 0.0001$; Cohen's $d = 1.47$, standardized mean difference), and this association holds under every sensitivity analysis.
   - **Underused Approaches:** Direct extraction, used in only 11 papers (10.1%), has the highest deployment rate (54.5%), whereas post-hoc methods (45 papers, 41.3%) and hybrid methods (11 papers, 10.1%) reach 15.6% and 18.2%, respectively. No paper produces fuzzy rules.
3. **Objective 3 (Network Security Relevance)** reveals:
   - **Network Relevance:** Network security relevance is high (mean Network_Score 4.12/5) but is not associated with deployment readiness ($r = -0.088$, $p = 0.3622$).
   - **DDoS Coverage:** Only 16 papers (14.7%) focus on DDoS as their primary research problem, and only 2 of them are deployment-ready.
   - **DDoS Validation and Adaptation:** 14 of the 16 DDoS-specific papers validate along at most one transferability dimension, and none of them includes an adaptive update mechanism.
4. **Objective 4 (Methodological Gaps and Limitations)** identified systematic limitations through Method_Score assessment, transferability analysis, cross-category analysis, and thematic synthesis:
   - **Methodological Rigor:** Methodological rigor is the highest-scoring component (mean Method_Score 4.27/5) and is not associated with deployment readiness ($r = 0.097$, $p = 0.3165$), so the gaps below are not explained by weak research design.
   - **Transferability Gap:** 33 papers (30.3%) report no transferability validation, 55 (50.5%) validate along a single dimension, and only 5 (4.6%) validate across all three dimensions (dataset, network environment, and attack type). None of the deployment-ready papers is validated across all three.
   - **Lack of Update Mechanisms:** Only 1 paper (0.9%) includes a mechanism for updating its deployed rules.
   - **Research Gap Themes:** this systematic review identified five recurring themes: the XAI-to-rule translation gap (59 papers, 54.1%), rule adaptation and robustness (22 papers, 20.2%), methodological issues (21 papers, 19.3%), cross-domain validation absence (20 papers, 18.3%), and computational overhead and scalability (16 papers, 14.7%). The translation gap is under-reported: only 44 of the 94 non-deployable papers (46.8%) acknowledge it.
   - **Persistent Gap over Time:** The share of papers without extractable rules rises from 13.0% in 2022–2023 to 70.4% in 2026, while publication year shows no significant association with deployment readiness ($r = -0.045$, $p = 0.6418$).

5.2 Literature Review Research Contributions

This review makes theoretical contributions to the intersection of explainable AI and network security:

- A four-category quality scoring system for XAI-guided rule extraction research. Among the reviews compared in Section 2.5, it is the only rubric that scores rule extraction quality as a separate dimension and uses the dimension scores in inferential analysis, linking study quality to practical deployment.
- Study-level statistical evidence for the rule translation gap: rule extraction quality, not XAI quality or methodological rigor, is the scoring dimension associated with deployment readiness, which quantifies the gap between explainability research and operational deployment requirements.
- Per-study classification of deployment readiness and of transferability along three dimensions (dataset, network environment, and attack type), together with a count of adaptive update mechanisms in XAI-guided rule extraction research (1 of 109 papers).
- Practical contributions for security practitioners by identifying why current XAI approaches fail to produce operational security rules.

5.3 Recommendations for Future Research

Based on this analysis, three research directions are proposed to address the identified gaps:

- **Creating automated rule extraction methods and workflows for XAI-IDS.** develop systematic translation methods that convert XAI outputs into deployment-ready rule formats, such as IF-THEN rules or rule-engine signatures, while maintaining interpretability and operational effectiveness. Direct extraction and hybrid pipelines show that this translation is feasible [36, 55].
- **Designing adaptive rule frameworks rather than static pattern matching.** Only one paper updates its deployed rules as conditions change [57]. Future work should create mechanisms for incremental rule updates as network environments evolve or attack patterns change, and should evaluate how rule quality is maintained after each update.
- **Establishing comprehensive evaluation standards** Future studies should validate rules across all three transferability dimensions (dataset, network environment, and attack type), including experiments in which a model trained on one dataset is tested on another, report rule-level metrics such as accuracy and false-positive rates on realistic benign traffic, then demonstrate deployment in simulated or realistic network environments.

Declarations

Funding The first author's doctoral studies are supported by a scholarship from the Ministry of Religious Affairs (MoRA) of the Republic of Indonesia and the Indonesia Endowment Fund for Education (LPDP) of the Ministry of Finance of the Republic of Indonesia (awardee number: BU05-231-0000302). No other specific grant was received for this research. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests Two of the authors (Mandahadi Kusuma and Mohd Yazid Idris) are authors of one of the studies included in this review [36]. To limit potential bias, that study was screened against the same eligibility criteria and scored with the same rubric as all other included studies, and excluding it does not change the main statistical findings (Section 3.1.3). The authors declare no other competing interests.

Ethics approval Not applicable. This article is a systematic review of published literature and did not involve human participants, animals, or personal data.

Consent to participate Not applicable.

Consent for publication Not applicable.

Data availability All data generated or analysed during this study are included in this published article and its supplementary information files. Supplementary File 1 contains the list of the 309 papers that reached the full-text stage, with the four-category scores of the 288 papers whose full texts were accessible. Supplementary File 2 contains the study-level data for the 109 included papers with the statistical summary computed from these data. Supplementary File 3 contains the data underlying Figures 3 and 5.

Use of generative AI During the preparation of this manuscript, the authors used a large language model (Claude, Anthropic) to cross-check the statistics and scoring data calculations from the extracted papers and check consistency between the text, tables, data files, LaTeX source of the manuscript, and to to repair grammatical errors. Screening, eligibility decisions, quality scoring, and data extraction were performed by the authors. All outputs were verified by the authors against the systematic research data.

Acknowledgements The first author's doctoral studies are supported by a scholarship from the Ministry of Religious Affairs (MoRA) and the Indonesia Endowment Fund for Education (LPDP) of the Ministry of Finance of the Republic of Indonesia.

References

1. O. Arreche, T. Guntur, M. Abdallah, Applied Sciences **14**(10), 4170 (2024). DOI 10.3390/app14104170. URL https://www.mdpi.com/2076-3417/14/10/4170
2. H.T. Pai, Y.H. Kang, W.C. Chung, IEEE Transactions on Information Forensics and Security **19**, 10302 (2024). DOI 10.1109/TIFS.2024.3488967. URL https://ieeexplore.ieee.org/document/10740319
3. S.M. Lundberg, S.I. Lee, in *Advances in Neural Information Processing Systems*, vol. 2017-December (Neural information processing systems foundation, 2017), vol. 2017-December, pp. 4766–4775
4. M.T. Ribeiro, S. Singh, C. Guestrin, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016 (Association for Computing Machinery, 2016), vol. 13-17-August-2016, pp. 1135–1144. DOI 10.1145/2939672.2939778
5. A.M. Souto, A. Sobrinho, L. Chaves e Silva, L. Dias da Silva, A. Ceccarelli, A. Perkusich, **40**, 102107. DOI 10.1016/j.iot.2026.102107. URL https://www.sciencedirect.com/science/article/pii/S2542660526002374
6. R. Nandhini, A. Senthil, N. Shri, Sainthathi, A.K. Prakash, S. Subramanian R, in *2026 2nd International Conference on Computing, Communication and Green Engineering (CCGE)*. pp. 1–6. DOI 10.1109/CCGE67142.2026.11581876. URL https://ieeexplore.ieee.org/abstract/document/11581876
7. F.L. Becerra-Suarez, I. Fernández-Roman, M.G. Forero, Mathematics **12** (2024). DOI 10.3390/math12091294
8. D. Kumar, R.K. Pateriya, R.K. Gupta, V. Dehalwar, A. Sharma, in *Procedia Computer Science*, vol. 218 (Elsevier B.V., 2022), vol. 218, pp. 2420–2429. DOI 10.1016/j.procs.2023.01.217
9. M.S. Sawah, H. Elmannai, A.A. El-Bary, K. Lotfy, O.E. Sheta, Scientific Reports **15**, 19063 (2025). DOI 10.1038/s41598-025-03868-x
10. R. Zhang, M. Tong, L. Chen, J. Xue, W. Liu, F. Xie, in *Proceedings - 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2020* (Institute of Electrical and Electronics Engineers Inc., 2020), pp. 736–743. DOI 10.1109/TrustCom50675.2020.00101

11. X. Xu, Y. Lai, X. Zhang, X. Dong, IEEE Transactions on Industrial Informatics **20**(8), 10624 (2024). DOI 10.1109/TII.2024.3396348. URL https://ieeexplore.ieee.org/document/10530913
12. M. Basak, D.W. Kim, M.M. Han, G.Y. Shin, Applied Sciences **15**(9), 5002 (2025). DOI 10.3390/app15095002. URL https://www.mdpi.com/2076-3417/15/9/5002
13. O. Arreche, T.R. Guntur, J.W. Roberts, M. Abdallah, IEEE Access **12**, 23954 (2024). DOI 10.1109/access.2024.3365140
14. P. Barnard, L.A. Dasilva, N. Marchetti, Authorea Preprints (2024). DOI 10.36227/TECHRXIV.170775198.87587642/V1. URL https://www.techrxiv.org/doi/pdf/10.36227/techrxiv.170775198.87587642/v1
15. M.M. Hasan, M.U. Islam, J. Uddin, SN Computer Science **4**(3), 1 (2023). DOI 10.1007/S42979-023-01744-X/FIGURES/4. URL https://link.springer.com/article/10.1007/s42979-023-01744-x
16. R. Kalakoti, R. Vaarandi, H. Bahşi, S. Nõmm, in *International Conference on Information Systems Security and Privacy*, vol. 1 (Science and Technology Publications, Lda, 2025), vol. 1, pp. 47–58. DOI 10.5220/0013180700003899
17. J. Ables, N. Childers, W. Anderson, S. Mittal, S. Rahimi, I. Banicescu, M. Seale, arXiv preprint (2024)
18. C. Molnar, G. Casalicchio, B. Bischl, in *Communications in Computer and Information Science*, vol. 1323 (Springer Science and Business Media Deutschland GmbH, 2020), vol. 1323, pp. 417–431. DOI 10.1007/978-3-030-65965-3_28
19. A.K. Dey, G.P. Gupta, S.P. Sahu, Arabian Journal for Science and Engineering pp. 1–20 (2024). DOI 10.1007/s13369-024-09578-2. URL https://link.springer.com/article/10.1007/s13369-024-09578-2
20. G.A. Mills, D.K. Acquah, R.A. Sowah, Journal of Computer Networks and Communications **2024** (2024). DOI 10.1155/2024/5775671
21. F. Uccello, M. Pawlicki, S. DAntonio, R. Kozik, M. Choraś, in *Proceedings of the Second International Conference on Advances in Computing Research (ACR24)*, ed. by K. Daimi, A. Al Sadoon (Springer Nature Switzerland, 2024), pp. 244–255. DOI 10.1007/978-3-031-56950-0_21
22. S. Chen, J. Liu, C. Chen, S. Xie, Z. Cheng, in *2024 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA)* (2024), pp. 1197–1202. DOI 10.1109/ISPA63168.2024.00160. URL https://ieeexplore.ieee.org/abstract/document/10885129
23. S. AL, S. Sagiroglu, Engineering Applications of Artificial Intelligence **144** (2025). DOI 10.1016/j.engappai.2025.110145
24. A. Bashaiwth, H. Binsalleeh, B. AsSadhan, Applied Sciences 2023, Vol. 13, Page 8820 **13**(15), 8820 (2023). DOI 10.3390/APP13158820. URL https://www.mdpi.com/2076-3417/13/15/8820/htmhttps://www.mdpi.com/2076-3417/13/15/8820
25. K. Scarfone, M. Souppaya, A. Cody, A. Orebaugh, Technical Guide to Information Security Testing and Assessment. Tech. Rep. NIST Special Publication (SP) 800-115, National Institute of Standards and Technology (2008). DOI 10.6028/NIST.SP.800-115. URL https://csrc.nist.gov/publications/detail/sp/800-115/final
26. M. Chakraborty, Archives of Computational Methods in Engineering **31**, 3535 (2024). DOI 10.1007/s11831-024-10089-4
27. A. Coscia, V. Dentamaro, S. Galantucci, A. Maci, G. Pirlo, Journal of Information Security and Applications **82** (2024). DOI 10.1016/j.jisa.2024.103736
28. J. Wang, C. Si, Z. Wang, Q. Fu, Computers, Materials and Continua **79**, 4297 (2024). DOI 10.32604/cmc.2024.050223
29. A. Tiwari, D. Kumar, International Journal of Engineering, Transactions B: Applications **38**, 1922 (2025). DOI 10.5829/IJE.2025.38.08B.16
30. Q. Liu, V. Hagenmeyer, H.B. Keller, **9**, 57542. DOI 10.1109/ACCESS.2021.3071263. URL https://ieeexplore.ieee.org/abstract/document/9395457
31. I.H. Sarker, H. Janicke, M.A. Ferrag, A. Abuadbba, **25**, 101110. DOI 10.1016/j.iot.2024.101110. URL https://www.sciencedirect.com/science/article/pii/S2542660524000520
32. V.Z. Mohale, I.C. Obagbuwa, **8**. DOI 10.3389/frai.2025.1526221. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1526221/full
33. A. Khediri, H. Slimi, A. Djeddar, H. Bendjenna, A. Yahiaoui, M. Derdour, in *2025 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI)*. pp. 1–4. DOI 10.1109/ICRAMI64946.2025.11472758. URL https://ieeexplore.ieee.org/abstract/document/11472758
34. H. Fasla, A. Mehmood, **59**(10), 201. DOI 10.1007/s10462-026-11581-y. URL https://doi.org/10.1007/s10462-026-11581-y
35. M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, D. Moher, BMJ **372**, n71 (2021). DOI 10.1136/bmj.n71. URL https://www.bmj.com/content/372/bmj.n71
36. M. Kusuma, M.Y. Idris, in *Proceedings of the 2026 18th International Conference on Computer Research and Development* (Association for Computing Machinery), ICCRD '26, pp. 184–191. DOI 10.1145/3796263.3796289. URL https://dl.acm.org/doi/10.1145/3796263.3796289
37. A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Information Fusion **58**, 82 (2020). DOI 10.1016/j.inffus.2019.12.012. URL https://www.sciencedirect.com/science/article/pii/S1566253519308103
38. B. Kitchenham, S. Charters, Technical Report **2** (2007)
39. R. Andrews, J. Diederich, A.B. Tickle, Knowledge-Based Systems **8**(6), 373 (1995). DOI 10.1016/0950-7051(96)81920-4. URL https://www.sciencedirect.com/science/article/pii/0950705196819204
40. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. (Routledge). DOI 10.4324/9780203771587
41. M.M. Shtayat, M.K. Hasan, R. Sulaiman, S. Islam, A.U.R. Khan, IEEE Access **11**, 115047 (2023). DOI 10.1109/ACCESS.2023.3323573. URL https://ieeexplore.ieee.org/document/10274903
42. S. Baral, S. Saha, A. Haque, in *2024 IEEE 10th World Forum on Internet of Things (WF-IoT)*. pp. 469–474. DOI 10.1109/WF-IoT62078.2024.10811456. URL https://ieeexplore.ieee.org/abstract/document/10811456
43. H. Zouhri, A. Idri, Future Generation Computer Systems **173**, 107882 (2025). DOI 10.1016/j.future.2025.107882. URL https://www.sciencedirect.com/science/article/pii/S0167739X25001773
44. M. Saied, S. Guirguis, Scientific Reports **15**(1), 7632 (2025). DOI 10.1038/s41598-025-90420-6. URL https://www.nature.com/articles/s41598-025-90420-6
45. O. Arreche, I. Bibers, M. Abdallah, IEEE Access **12**, 83830 (2024). DOI 10.1109/ACCESS.2024.3407029. URL https://ieeexplore.ieee.org/document/10540382
46. M. Sarhan, S. Layeghy, M. Portmann, Big Data Research **30**, 100359 (2022). DOI 10.1016/j.bdr.2022.100359. URL https://www.sciencedirect.com/science/article/pii/S2214579622000533
47. T. Tahrim, M.A. Sharan, M.F. Ahmed, A. Shakil, in *2024 27th International Conference on Computer and Information Technology (ICCIT)*. pp. 1392–1397. DOI 10.1109/ICCIT64611.2024.11022481. URL https://ieeexplore.ieee.org/abstract/document/11022481
48. H. Lee, T. Kwon, J. Lee, J. Song, in *2024 IEEE Conference on Dependable and Secure Computing (DSC)* (2024), pp. 147–154. DOI 10.1109/DSC63325.2024.00039. URL https://ieeexplore.ieee.org/abstract/document/10763545
49. E. Namjoo, A.N. O'Connor, J. Buckley, C. Ryan, **15**(12), 6931. DOI 10.3390/app15126931. URL https://www.mdpi.com/

2076-3417/15/12/6931
50. B. Hsupeng, K.W. Lee, T.E. Wei, S.H. Wang, in *2022 24th International Conference on Advanced Communication Technology (ICACT)* (2022), pp. 27–33. DOI 10.23919/ICACT53585.2022.9728897. URL https://ieeexplore.ieee.org/document/9728897
51. S. Bayrak, **66**(11), 6675. DOI 10.1007/s10115-024-02181-9. URL https://link.springer.com/article/10.1007/s10115-024-02181-9
52. D.L. Marino, C.S. Wickramasinghe, C. Rieger, M. Manic, IEEE Transactions on Industrial Informatics **21**(5), 4252 (2025). DOI 10.1109/TII.2025.3534443. URL https://ieeexplore.ieee.org/document/10906675
53. Y. Chen, J. Li, N. Guo, **22**(3), 679. DOI 10.1007/s10207-022-00656-w. URL https://link.springer.com/article/10.1007/s10207-022-00656-w
54. A. Okutan Kara, A. Boyacı, **16**(12), 5912. DOI 10.3390/app16125912. URL https://www.mdpi.com/2076-3417/16/12/5912
55. M. Wichtlhuber, E. Strehle, D. Kopp, L. Prepens, S. Stegmueller, A. Rubina, C. Dietzel, O. Hohlfeld, in *Proceedings of the ACM SIGCOMM 2022 Conference* (Association for Computing Machinery), SIGCOMM '22, pp. 707–722. DOI 10.1145/3544216.3544268. URL https://dl.acm.org/doi/10.1145/3544216.3544268
56. J. Piet, D. Nwoji, V. Paxson, in *Proceedings of the ACM SIGCOMM 2023 Conference* (Association for Computing Machinery, 2023), ACM SIGCOMM '23, pp. 850–866. DOI 10.1145/3603269.3604840. URL https://dl.acm.org/doi/10.1145/3603269.3604840
57. M. Fan, J. Zuo, J. Zhu, Y. Lu, IEEE Internet of Things Journal **12**(12), 19504 (2025). DOI 10.1109/JIOT.2025.3542142. URL https://ieeexplore.ieee.org/abstract/document/10887229
58. D. Han, Z. Wang, R. Feng, M. Jin, W. Chen, K. Wang, S. Wang, J. Yang, X. Shi, X. Yin, Y. Liu, in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (Association for Computing Machinery, 2024), CCS '24, pp. 4509–4523. DOI 10.1145/3658644.3670375. URL https://dl.acm.org/doi/10.1145/3658644.3670375
59. P. Toralkar, K. Mainalli, S. Allagi, S.K. Debnath, S. Bagchi, W.Y. Leong, M.N.A. Khan, **6**(4), 381. DOI 10.1007/s42979-025-03859-9. URL https://doi.org/10.1007/s42979-025-03859-9
60. W. Li, D. Ma, Z. Li, H. Bao, S. Wang, H. Jin, X.Y. Zhang, in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (Association for Computing Machinery), CCS '24, pp. 4934–4936. DOI 10.1145/3658644.3691410. URL https://dl.acm.org/doi/10.1145/3658644.3691410
61. F. Uccello, M. Pawlicki, S. D'Antonio, R. Kozik, M. Choraś, in *Computational Science and Its Applications – ICCSA 2024 Workshops*, ed. by O. Gervasi, B. Murgante, C. Garau, D. Taniar, A.M.A. C. Rocha, M.N. Faginas Lago (Springer Nature Switzerland, 2024), pp. 100–112. DOI 10.1007/978-3-031-65223-3_7
62. S. Verma, S. Prabakeran, in *2025 8th International Conference on Trends in Electronics and Informatics (ICOEI)* (2025), pp. 1254–1261. DOI 10.1109/ICOEI65986.2025.11013553. URL https://ieeexplore.ieee.org/document/11013553
63. K. Dietz, M. Hajizadeh, J. Schleicher, N. Wehner, S. Geißler, P. Casas, M. Seufert, T. Hoßfeld, in *2024 20th International Conference on Network and Service Management (CNSM)*. pp. 1–7. DOI 10.23919/CNSM62983.2024.10814448. URL https://ieeexplore.ieee.org/abstract/document/10814448
64. D. Trizna, L. Demetrio, B. Biggio, F. Roli, IEEE Transactions on Information Forensics and Security **19**, 6155 (2024). DOI 10.1109/TIFS.2024.3409083. URL https://ieeexplore.ieee.org/abstract/document/10551436
65. P. Patel, D. Shah, D. Shah, F. Ramoliya, R. Gupta, S. Tanwar, in *2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)* (2025), pp. 598–604. DOI 10.1109/IC3ECSBHI63591.2025.10991074. URL https://ieeexplore.ieee.org/abstract/document/10991074
66. N. Ayoughi, S. Nejati, M. Sabetzadeh, P. Saavedra, in *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems* (Association for Computing Machinery, 2024), MODELS '24, pp. 172–182. DOI 10.1145/3640310.3674087. URL https://doi.org/10.1145/3640310.3674087
67. C. Minh, K. Vermeulen, C. Lefebvre, P. Owezarski, W. Ritchie, in *2023 19th International Conference on Network and Service Management (CNSM)* (2023), pp. 1–9. DOI 10.23919/CNSM59352.2023.10327818. URL https://ieeexplore.ieee.org/abstract/document/10327818
68. L. Shafir, R. Giryes, A. Wool, **34**, 1205. DOI 10.1109/TON.2025.3617580. URL https://ieeexplore.ieee.org/abstract/document/11201273
69. M.F. Saiyed, I. Al-Anbagi, **13**(7), 12958. DOI 10.1109/JIOT.2026.3653260. URL https://ieeexplore.ieee.org/abstract/document/11346516
70. J. Wilson, A.P. Deshpande, M. Premkumar, T. Yuvaraja, **33**(3), 1056. DOI 10.17559/TV-20250720002849. URL https://hrcak.srce.hr/clanak/500458
71. O. Ekpo, V. Casola, P. Asuquo, B. Agbor, in *2026 IEEE 4th International Conference on Mobility, Operations, Services and Technologies (MOST)*. pp. 147–159. DOI 10.1109/MOST69733.2026.00025. URL https://ieeexplore.ieee.org/abstract/document/11527449
72. F.N. Dresch, F.H. Scherer, M.M. Ciocca, V.E. Quincozes, S.E. Quincozes, D. Kreutz, **32**(1), 733. DOI 10.5753/jbcs.2026.5664. URL https://journals-sol.sbc.org.br/index.php/jbcs/article/view/5664
73. M.O. Onuorah, Y. Sun, D. Mashao, **14**, 46085. DOI 10.1109/ACCESS.2026.3675887. URL https://ieeexplore.ieee.org/abstract/document/11443277
74. R.G. Garroppo, P.G. Giardina, G. Landi, M. Ruta, **17**(5), 191. DOI 10.3390/fi17050191. URL https://www.mdpi.com/1999-5903/17/5/191
75. F. Uccello, S. Nadjm-Tehrani, **99**, 104448. DOI 10.1016/j.jisa.2026.104448. URL https://www.sciencedirect.com/science/article/pii/S2214212626000785
76. B.G. Paltun, R. Fuladi, R. El Malki, in *2025 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. pp. 145–150. DOI 10.1109/EuCNC/6GSummit63408.2025.11037121. URL https://ieeexplore.ieee.org/document/11037121
77. E. Antony, A.K. Roy, **286**, 112399. DOI 10.1016/j.comnet.2026.112399. URL https://www.sciencedirect.com/science/article/pii/S1389128626004111
78. S. Niktabe, D. Prasad Sharma, A. Habibi Lashkari, **82**(5), 352. DOI 10.1007/s11227-026-08494-6. URL https://doi.org/10.1007/s11227-026-08494-6
79. D. Supriya, A.V. Praveen Krishna, **8**, 200459. DOI 10.1016/j.sasc.2026.200459. URL https://www.sciencedirect.com/science/article/pii/S2772941926000232
80. E. Arkadopoulou, N. Kostopoulos, M. Grammatikou, V. Maglaris, **14**, 64220. DOI 10.1109/ACCESS.2026.3687021. URL https://ieeexplore.ieee.org/abstract/document/11493868
81. M. Marwan, M. Chemmakha, A. Ait Temghart, M. Lazaar, **6**(1), 255. DOI 10.1007/s44163-026-00955-4. URL https://doi.org/10.1007/s44163-026-00955-4
82. J. Ding, P. Shan, R. Ding, E.Y. Zhu, H. Li, Y. Wu, T. Kang, Y. Qiao, C. Jiang, in *SoutheastCon 2026*. pp. 1–7. DOI 10.1109/SoutheastCon63549.2026.11476109. URL https://ieeexplore.ieee.org/abstract/document/11476109
83. G. Qian, W. He, **9**(1), 32. DOI 10.1186/s42400-025-00438-6. URL https://doi.org/10.1186/s42400-025-00438-6

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Datafiguretrends.xlsx
- scoringincludedexcluded309papers.xlsx
- 109includedpapersmasterfiles.xlsx