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.

Towards Sociotechnical Intelligence in Cyber Intrusion Detection: A Systematic Review Integrating ML/DL Performance, Human Adoption, and RAG-Enhanced Explainable AI

Yali Ren
Georgia Institute of Technology
Ning Wang
nwang@uh.edu
University of Houston

Systematic Review

Keywords: Intrusion Detection Systems (IDS), Distributed and Federated Learning, RAG‐Augmented Explainable AI, Edge–Cloud Security Architecture, Performance–Adoption Gap (PAG), Sociotechnical Systems

Posted Date: April 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9329228/v1

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

Additional Declarations: The authors declare no competing interests.

Towards Sociotechnical Intelligence in Cyber Intrusion Detection: A Systematic Review Integrating ML/DL Performance, Human Adoption, and RAG-Enhanced Explainable AI

Abstract

Intelligent Intrusion Detection Systems (IDS) routinely report accuracy above 99% on standard benchmarks, yet real-world deployment remains limited, an operational divergence we formalize as the Performance–Adoption Gap (PAG). This Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) compliant systematic survey integrates 153 empirical IDS studies (2009–2025), an extended UTAUT2 adoption survey of 300 security professionals, and the distributed-systems literature to examine this gap from a sociotechnical perspective. We contribute: (1) a transparent six-stage PRISMA pipeline with reproducibility artifacts; (2) a Six-Axis Sociotechnical Taxonomy that extends four technical IDS dimensions with two empirically grounded human-factors axes; (3) a quantitative meta-analysis showing a mean accuracy of 97.1% but a median minority-class recall of 0.648 and an average cross-dataset accuracy drop of 18 percentage points; (4) a formal definition of the PAG as a divergence measure; and (5) a ten-stage Sociotechnical IDS Architecture that maps detection, explanation, and analyst-interaction stages to distributed edge–cloud and federated deployment scenarios. The analysis highlights scalability constraints, including inference latency, horizontal partitioning, and model-synchronization overhead, that directly affect cluster-based IDS deployments. Finally, we outline a prioritized research roadmap, identifying Retrieval-Augmented Generation augmented (RAG-augmented) explainability as a high-leverage direction for bridging technical performance with operational adoption in distributed and federated IDS environments.

Keywords: Intrusion Detection Systems (IDS), Distributed and Federated Learning, RAG-Augmented Explainable AI, Edge–Cloud Security Architecture, Performance–Adoption Gap (PAG), Sociotechnical Systems

1 Introduction

The cybersecurity landscape confronts a mounting operational crisis. Cybercrime costs are projected to reach $10.5 trillion annually by 2025 [1], a 3.5 million global shortfall

in security professionals persists [2], and Security Operations Center (SOC) analysts process more than 11,000 alerts per day [3], the majority of which are false positives that erode analyst effectiveness. Simultaneously, the attack surface expands through cloud migrations, Internet of Things (IoT) proliferation, and Artificial Intelligence (AI)-enabled adversaries deploying polymorphic malware and adaptive evasion techniques [4, 5].

Machine learning (ML) and deep learning (DL), hereafter ML/DL, have substantially advanced IDS over the past decade: DL models account for 40% of recent IDS publications, ensemble approaches for 35%, and emerging frameworks, transformers, Graph Neural Networks (GNNs), federated learning, and large language model (LLM)-based detection, for another 22%. Benchmark accuracy regularly exceeds 98% on widely used datasets. Yet technically superior systems are rarely deployed in production. Industry surveys indicate that while 87% of security leaders recognise AI's potential value, only 32% have successfully deployed generative AI tools in security operations [6], a paradox motivating this review.

Two complementary research traditions address different dimensions of this challenge but have evolved in near-complete isolation. The technical IDS literature catalogues models, preprocessing pipelines, and evaluation metrics. The human-factors literature applies technology acceptance frameworks, notably UTAUT2 [23], to measure why security professionals adopt or reject AI tools [4], consistently identifying privacy concerns, trust in AI outputs, and risk perception as adoption determinants at least as influential as technical performance. Neither stream routinely measures the other's primary outcomes.

The contribution of this paper is to bridge these streams through a comprehensive, PRISMA-compliant systematic review. We do not claim to have invented the sociotechnical framing, important precedents exist in usable security [7], human-AI interaction [8], and earlier IDS surveys [9, 10]. Rather, we operationalise it with quantitative precision, extend it with a formal gap measure, and connect it to practical distributed deployment considerations relevant to the Cluster Computing context.

Paper organisation. Section 2 situates this survey relative to prior work. Section 3 describes methodology. Section 4 presents benchmark datasets, taxonomy, technical paradigms, and meta-analysis. Section 5 presents the PAG formalisation, Sociotechnical Architecture, and Dual-Metric Framework. Section 6 covers the Future Research Roadmap and validity discussion. Section 7 concludes.

2 Related Work

2.1 Technical IDS Surveys

Several prior surveys cover the technical landscape of ML/DL-based intrusion detection. Liao et al. [10] provided an early taxonomy of IDS techniques, distinguishing anomaly-, signature-, and specification-based approaches. Buczak and Guven [9] surveyed ML methods for cyber intrusion detection, covering feature selection, classifier comparison, and dataset evaluation, but predated the DL dominance now visible in our corpus. Ferrag et al. [11] reviewed DL for cyber-intrusion detection across IoT and cloud environments, cataloguing Convolutional Neural Networks (CNNs), Long

Short-Term Memory (LSTM) networks, Generative Adversarial Networks (GANs), and autoencoder approaches on then-available benchmarks. Abdulganiyu et al. [12] offered an updated systematic review focused on DL architectures. Collectively, these surveys establish the technical foundation; they do not address the human adoption dimension, do not formalise a performance–adoption divergence, and predate LLM/RAG-based approaches.

Our survey extends this line by: (a) incorporating human-factors axes systematically absent from prior technical reviews; (b) covering the 2009–2025 period including transformer, GNN, federated, and LLM/RAG paradigms; and (c) quantifying cross-dataset generalisation failure and minority-class recall collapse with confidence intervals.

2.2 XAI in Cybersecurity

Arrieta et al. [13] provided a landmark survey of Explainable Artificial Intelligence (XAI) across domains, introducing a taxonomy of intrinsic, post-hoc local, and post-hoc global methods applicable to security. Lundberg and Lee [14] introduced SHapley Additive exPlanations (SHAP) values, now the dominant post-hoc explainability tool in IDS research. Marino et al. [15] and Arreche et al. [16] specifically evaluated XAI for network intrusion detection, identifying that SHAP feature scores lack actionability for SOC analysts without domain translation.

Retrieval-Augmented Generation (RAG) architectures for knowledge-grounded generation were introduced by Lewis et al. [17] and subsequently applied to cybersecurity by Houssel et al. [18] and Blefari et al. [19]. The specific combination of SHAP attribution with RAG-retrieved threat intelligence narratives, which we refer to as RAG-XAI, appears as a design proposal in these recent works but has not yet been systematically evaluated in peer-reviewed IDS studies within our corpus. We therefore characterise it as a *promising but unvalidated* intervention, not a proven solution, and identify its empirical evaluation as a priority research direction.

2.3 Human-Centered Security and Technology Adoption

The usable security tradition [7, 20] established that security systems must be designed for human operators, not solely for technical adversaries. Bada et al. [21] reviewed psychological and behavioural factors in cybersecurity adoption. The Unified Theory of Acceptance and Use of Technology (UTAUT) framework [22] and its extension UTAUT2 [23] are the dominant models for predicting technology adoption intentions; Ren [4] applied an extended UTAUT2 incorporating RAG-specific trust constructs to AI cybersecurity assistants.

Transparency note on Reference [4]. The UTAUT2 study [4] was authored by the first author of the present survey and published in a venue (Machine Learning Research, Science Publishing Group) that appears on some lists of publishers with limited peer-review oversight. We acknowledge this limitation explicitly. The present survey relies on the UTAUT2 findings as one of two empirical pillars, but treats them as *preliminary pending independent replication*. The survey instrument, anonymised

dataset, and full CFA output are available from the corresponding author to facilitate such replication. Readers should weigh the adoption findings accordingly, and we identify independent replication of the UTAUT2 study as a priority (Section 6).

Lee and See [24] established that optimal human-automation performance requires calibrated trust.

What distinguishes our survey from these human-factors studies is the systematic linkage to a quantitative IDS technical evidence base: we do not study adoption in isolation but compare adoption barriers directly against measurable technical deficiencies catalogued in 153 empirical papers.

2.4 Sociotechnical Systems Framing

Sociotechnical systems theory, tracing to the Tavistock Institute [25] and formalised for software systems by Baxter and Sommerville [26], holds that systems must be designed for the joint optimisation of technical and social subsystems. Amershi et al. [8] operationalised this for AI systems with 18 design guidelines. We apply and extend this framing specifically to IDS, contributing a taxonomy that operationalises sociotechnical design dimensions as codeable, measurable axes, a step not taken in prior sociotechnical security literature.

2.5 Positioning of This Survey

Table 1 summarises how this survey relates to the closest prior works. Each prior survey made significant contributions within its scope: Liao et al. established foundational IDS taxonomies that remain widely adopted; Buczak and Guven provided the most comprehensive early ML-for-IDS feature-selection analysis; Ferrag et al. pioneered systematic DL architecture comparison across IoT environments; Abdulganiyu et al. offered the most recent DL-focused systematic review with quality assessment; and Arrieta et al. produced the landmark cross-domain XAI taxonomy. The distinguishing contribution of the present survey is the *integration* of quantitative meta-analysis, formalised gap measurement, adoption modelling, and distributed deployment analysis within a single PRISMA-compliant framework, not a claim that prior surveys are deficient in areas outside their stated scope.

Table 1 Comparison of this survey with closely related prior works. CIs = confidence intervals; Distrib. = distributed systems coverage.

Survey / Study	PRISMA	Meta-analysis	XAI	Human factors	LLM/RAG	Distrib.
Liao et al. (2013) [10]	No	No	No	No	No	No
Buczak & Guven (2016) [9]	No	Partial	No	No	No	No
Ferrag et al. (2020) [11]	No	No	No	No	No	Partial
Abdulganiyu et al. (2023) [12]	Partial	No	No	No	No	No
Arrieta et al. (2020) [13]	No	No	Yes	No	No	No
Ren (2026) [4]	No	No	No	Yes	Partial	No
This survey	**Yes**	**Yes (+CIs)**	**Yes**	**Yes**	**Yes**	**Yes**

3 Method

3.1 Protocol and Search Strategy

This review follows Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [27]. Five databases were searched: IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and arXiv (cs.CR and cs.LG categories). Boolean search strings:

- **Primary:** ("intrusion detection" OR "IDS" OR "network intrusion") AND ("machine learning" OR "deep learning" OR "neural network" OR "data science")
- **Secondary:** ("anomaly detection" OR "cyber-attack detection") AND ("classification" OR "ensemble" OR "federated learning" OR "explainable AI")

Time range: January 2009–December 2025; language: English only.

Scope delimitation. This survey focuses on *network-based* intrusion detection systems, that is, systems that analyse network traffic (packet captures, NetFlow/IPFIX records, or aggregated flow features) to identify malicious activity. Host-based IDS (HIDS), application-layer detection, and Endpoint Detection and Response (EDR) systems are outside the scope of this review. While these components are critical in modern SOC infrastructure and share some ML/DL techniques with network IDS, they operate on fundamentally different data modalities (system calls, process telemetry, file-system events) and face distinct deployment constraints. Readers interested in host-based or endpoint detection are referred to specialised surveys in those domains.

3.2 Inclusion and Exclusion Criteria

Papers were included if they: (1) proposed or evaluated an ML/DL/data-science (DS)-based IDS or component; (2) provided quantitative evaluation on a named dataset with performance metrics; (3) appeared in a peer-reviewed venue, *or* in an arXiv preprint with $\geq$10 citations *and* publicly available code (the $\geq$10 citation threshold was chosen to exclude unpublished drafts without peer exposure; we acknowledge that this threshold introduces a recency bias against high-quality 2024–2025 preprints that have not yet accumulated citations, the sensitivity analysis in Section 6.2 demonstrates that excluding all arXiv papers changes aggregate results by $<$0.3 pp, confirming that our conclusions do not depend on arXiv-sourced papers; however, we retain them because three arXiv-sourced papers contribute unique coverage of LLM/RAG-based IDS approaches not yet appearing in peer-reviewed venues, and their removal would create a gap in the emerging-paradigm analysis); and (4) were fully accessible in English. Papers were excluded if they were survey/review papers without novel empirical contribution, non-network security applications (spam, fraud), lacked quantitative evaluation, or were from predatory/unverifiable venues identified via Beall's list.

3.3 Paper Selection Pipeline (PRISMA Flow)

Table 2 details the six-stage selection pipeline. Inter-rater reliability was assessed on a 30-paper random sample by two independent coders (Cohen's $\kappa = 0.79$, substantial

agreement); disagreements were resolved by structured discussion. For human-factors axes (A1–A2), coders recorded whether adoption outcomes were measured and whether user studies or deployment evaluations were reported.

Table 2 PRISMA paper selection pipeline

Stage	Records	Retained	Exclusion notes
Database search	1,847	1,847	Boolean search; Jan 2009–Dec 2025
Deduplication	1,847	1,412	435 duplicates (DOI + title matching)
Title + Abstract screening	1,412	487	Not IDS-focused (n = 412); no ML/DL (n = 298); not empirical (n = 215)
Full-text eligibility	487	203	No quantitative eval (n = 98); insufficient detail (n = 86); non-English (n = 42); predatory (n = 58)
Quality scoring ($\geq$3/5)	203	153	Dataset unspecified (n = 22); no baseline (n = 18); not reproducible (n = 10)
Final corpus	**153**	**153**	91% peer-reviewed; 9% arXiv ($\geq$10 citations, code available)

4 Results

4.1 Benchmark Dataset Landscape

Dataset quality fundamentally shapes IDS research trajectories. NSL-KDD [52] (2009, 148K records) remains a widely used historical baseline despite known statistical artefacts. UNSW-NB15 [28] (2015, 2.54M records, 9 attack types, 49 features) and CIC-IDS2017 [29] (2.83M records, 7 attack types) together account for 71% of dataset usage in our corpus, producing a benchmark monoculture that, as the cross-dataset analysis demonstrates (Section 4.9), conceals generalisation failures.

IoT-specific benchmarks have grown rapidly: Bot-IoT [30] (72M+ records), TON_IoT [31] (461K records), CIC-IoT2023 [32] (46M+ records, 33 attack scenarios), Edge-IIoTset [33] (20M+ records, industrial IoT), and CICIoMT2024 [34] (40 medical devices, 18 attack types). IoT datasets appear in 28% of papers, sharply increasing since 2022, consistent with the expanding distributed edge-computing attack surface. Fewer than 8% of studies use real-world or proprietary traffic, a recognised limitation that inflates reported benchmark performance.

4.2 Six-Axis Sociotechnical Taxonomy

The Six-Axis Sociotechnical Taxonomy classifies IDS research along six dimensions (Table 3). Axes T1–T4 are grounded in established technical literature; Axes A1–A2 extend prior work by systematically operationalising adoption-readiness and sociotechnical fit as codeable design properties, an extension not present in earlier IDS taxonomies [9, 10].

Important clarification on axes A1–A2. A key empirical finding from the coding exercise is that *none* of the 153 technical IDS papers measure either A1 or A2 outcomes. This means that, unlike T1–T4 which classify observed variance across

the corpus, A1 and A2 have *zero variance* and therefore cannot function as taxonomy axes for classifying existing work. We retain them in the taxonomy framework for a different purpose: they function as *aspirational design requirements*, dimensions along which future IDS research *should* vary but does not yet. Their inclusion is motivated by the UTAUT2 adoption evidence [4], which identifies measurable adoption determinants (Privacy Concerns $\beta = -0.246$; Trust $\beta = +0.223$) that map onto A1–A2 categories. We explicitly acknowledge that A1 and A2 are derived from a single cross-sectional survey study rather than from multi-study empirical validation, and their independence from each other has not been tested through discriminant analysis. Future work should validate whether these represent empirically distinct and exhaustive dimensions of adoption readiness.

4.3 Classical Machine Learning Methods

Classical ML methods, logistic regression, Naïve Bayes, decision trees, k-Nearest Neighbour (k-NN), Support Vector Machine (SVM), random forest, and gradient boosting, appear in 53% of reviewed papers, primarily as baselines. Mishra et al. [35] reported that preprocessing quality and feature selection strategy produce greater performance variance than classifier choice, a finding corroborated by Alhajjar et al. [53], who demonstrated that feature engineering choices dominate accuracy differences across classifiers on NSL-KDD, by Sarhan et al. [54], who showed that standardising feature sets across datasets narrows inter-model performance gaps substantially, and by Ring et al. [55], whose cross-dataset evaluation confirmed that preprocessing pipeline design accounts for a larger share of performance variance than model architecture selection. This implies that the community's emphasis on novel architectures may be partially misallocated relative to pipeline design investment.

Random Forest achieves 97–99% accuracy on UNSW-NB15 through bootstrap aggregation, with native feature importance as post-hoc explainability [36]. XGBoost achieves state-of-the-art performance on tabular benchmarks [37]. Kilincer et al. [38] demonstrated dataset-dependent model rankings, cautioning against single-dataset conclusions. Classical ML in our corpus: aggregate accuracy 95–99%; minority-class recall 55–85%, systematically below aggregate figures.

4.4 Deep Learning Approaches

CNNs and Spatial Feature Extraction

CNNs apply convolutional filters to traffic feature vectors to capture local feature correlations, achieving 97–99% accuracy on CIC-IDS2017 and UNSW-NB15 [12]. Gradient-weighted Class Activation Mapping (Grad-CAM) provides visual attribution for image-encoded traffic but is less effective for raw tabular flow features, limiting SOC interpretability.

RNNs, LSTMs, and Temporal Modelling

Recurrent Neural Networks (RNNs), specifically LSTMs, model sequential dependencies in traffic flows, enabling detection of multi-step attack campaigns [39].

Table 3 Six-Axis Sociotechnical Taxonomy—Technical Axes (T1–T4): categories, corpus evidence, and design trade-offs. RL = Reinforcement Learning; FPR = False Positive Rate; SMOTE = Synthetic Minority Over-sampling Technique; ADASYN = Adaptive Synthetic Sampling; GAN = Generative Adversarial Network; SDN = Software-Defined Networking; CPS = Cyber-Physical Systems; SCADA = Supervisory Control and Data Acquisition.

T1: Learning Strategy	
Categories	**(a)** Supervised; **(b)** Unsupervised; **(c)** Semi-supervised; **(d)** Self-supervised; **(e)** Reinforcement Learning (RL).
Evidence	87% supervised; self-supervised growth post-2022 as a strategy to reduce label dependency; RL <3% owing to sample-efficiency challenges in high-dimensional traffic spaces.
Trade-offs	**Supervised:** high benchmark accuracy; requires labelled data; poor generalisation to unseen attacks. **Unsupervised/Semi:** zero-day capable; higher False Positive Rate (FPR). **Self-supervised:** reduces labelling cost; emerging. **RL:** adaptive; computationally intensive; fragile under adversarial reward hacking.
T2: Imbalance Handling	
Categories	**(a)** None / Implicit; **(b)** Sampling: SMOTE / ADASYN; **(c)** Cost-sensitive learning; **(d)** Ensemble-recall weighting; **(e)** Generative: GAN / VAE synthesis.
Evidence	62% employ some imbalance-handling; only 21% optimise per-class recall; GAN-based synthesis <8%. Recall hierarchy: recall-aware ensemble (0.94) > VAE synthesis (0.87) > ADASYN (0.79) > SMOTE (0.73), not yet reflected in method preferences.
Trade-offs	**SMOTE/ADASYN:** simple; may overfit synthetic minority boundary. **Cost-sensitive:** no data modification; effective for moderate imbalance. **Ensemble-recall:** best minority-class recall (0.94); increased complexity. **GAN/VAE:** most realistic; high training cost; GAN instability.
T3: Explainability Level	
Categories	**(a)** Intrinsic (rule-based, decision tree); **(b)** Post-hoc local: SHAP / LIME; **(c)** Post-hoc global (surrogate models); **(d)** Hybrid XAI (attention + SHAP); **(e)** RAG-augmented XAI (proposed).
Evidence	66% no or minimal XAI; SHAP or LIME in 28%; RAG-XAI <3% (emerging proposals, not yet peer-reviewed). SHAP scores unactionable for SOC analysts without domain translation [16]. RAG-XAI maps SHAP to CVE/NVD and MITRE ATT&CK narratives.
Trade-offs	**Intrinsic:** fully transparent; limited capacity; maps to SOC playbooks. **SHAP/LIME:** model-agnostic; LIME unstable; SHAP expensive at line rate. **RAG-augmented:** projected highest SOC usability (6.4/7); latency overhead; requires curated knowledge base; **empirical validation pending**.
T4: Deployment Context	
Categories	**(a)** Cloud / Data Centre (DC); **(b)** Edge / IoT; **(c)** Federated (multi-organisational); **(d)** Software-Defined Networking (SDN); **(e)** Industrial CPS / SCADA.
Evidence	Cloud-focused: 58%; IoT/Edge: 28% (growing since 2022); Federated: 6%; SDN: 5%; CPS/SCADA: <3%. Only 6% address federated or edge deployment; edge-cloud co-design remains an open scalability problem for cluster computing contexts.
Trade-offs	**Cloud/DC:** unlimited compute; centralisation risk. **Edge/IoT:** constrained hardware; $\leq$1 ms inference for line-rate detection. **Federated:** privacy-preserving; non-IID challenge; communication overhead. **SDN:** programmable traffic steering; vendor lock-in risk. **CPS/SCADA:** safety-critical; legacy protocol constraints.

CNN–LSTM hybrids deliver the best accuracy–efficiency trade-off in IoT benchmarks [40], combining spatial and temporal feature learning. Autoencoders detect zero-day attacks through reconstruction error; Variational Autoencoders (VAEs) improve generalisation by imposing distributional constraints on the latent space [41].

Table 4 Six-Axis Sociotechnical Taxonomy—Human-Factors Axes (A1–A2): categories, corpus evidence, and design trade-offs (continued from Table 3).

A1: Adoption Readiness	
Categories	**(a)** Privacy-compliant (on-premise / DP certified); **(b)** Trust-transparent (citation-visible explanations); **(c)** Risk-communicated (robustness certificates); **(d)** Explainability-verified (analyst-rated quality).
Evidence	0% of 153 technical IDS papers measure Adoption Readiness—confirming the PAG [4]. UTAUT2 evidence: Privacy Concerns ($\beta = -0.246$, $p < 0.001$) strongest negative predictor; Trust & Accuracy ($\beta = +0.223$, $p < 0.001$) strongest positive XAI predictor; Low Risk Belief ($\beta = +0.210$, $p < 0.001$) enables management approval.
Trade-offs	**Privacy compliance:** on-premise or DP certification overcomes $\beta = -0.246$ barrier. **Citation transparency:** RAG citations enable analyst verification, addressing Trust deficit. **Risk communication:** robustness certificates unlock Low Risk Belief adoption pathway. **Explainability verification:** analyst-panel ratings (7-point Likert) required for deployment claims.
A2: Sociotechnical Fit	
Categories	**(a)** Performance-primary (productivity gains demonstrated); **(b)** Effort-optimised (seamless SIEM integration); **(c)** Influence-sensitive (peer endorsement for novices); **(d)** Habit-forming (workflow integration).
Evidence	UTAUT2 base $R^2 = 0.412$; extended model $R^2 = 0.564$ ($\Delta R^2 = 0.152$, $p < 0.001$) [4]. Performance Expectancy ($\beta = +0.235$) strongest positive predictor; Effort Expectancy ($\beta = +0.156$) stronger for technical users ($\beta = 0.187$) vs. non-technical ($\beta = 0.098$); Social Influence non-significant overall but strong for novices ($\beta = 0.201$).
Trade-offs	**Performance-primary:** SOC productivity gains drive adoption more than accuracy metrics. **Effort-optimised:** one-click SIEM adapters and progressive disclosure UIs reduce perceived effort. **Influence-sensitive:** novices respond to peer endorsement; experienced analysts are performance-driven. **Habit-forming:** automated ATT&CK mapping within existing tools enables habitual use.

Transformers, GNNs, and Emerging Architectures

Transformer self-attention captures long-range flow dependencies [42]. In our corpus, transformer models constitute only 8% of papers but achieve the highest minority-class recall (80–90%), a critical advantage given the minority-attack collapse problem identified in Section 4.9. GNNs represent traffic as host-communication graphs, detecting lateral movement and coordinated scanning invisible to flat feature-vector analysis [43, 44]. GNNs appear in 5% of the corpus with minority-class recall 80–88%.

4.5 Federated Learning, Distributed Deployment, and Adversarial ML

Federated and Distributed IDS

Federated learning enables collaborative IDS training without centralising sensitive traffic data [45], growing from 0% pre-2019 to 6% of papers in 2023–2025. This growth is directly relevant to cluster computing deployments, where privacy-preserving distributed training across organisational boundaries is a primary design requirement. Three technical challenges limit current federated IDS: non-IID data distributions (addressed by pFedCross [46] through Jensen-Shannon divergence optimisation); adversarial participant-poisoning; and communication overhead in resource-constrained IoT environments. IBM X-Force (2024) reports that 60% of enterprise

breaches involve lateral movement across organisational boundaries, precisely the attack pattern distributed federated IDS is designed to detect.

Scalability and Edge-Cloud Coordination

Cluster computing deployments introduce three IDS-specific scalability constraints. First, *throughput*: line-rate detection at 100 Gbps demands inference latency below $\approx$1 μs per flow, achievable only by lightweight classifiers (Random Forest, XGBoost) or specialised hardware accelerators; transformer and GNN architectures require edge-cloud offloading of computationally intensive inference. An analytical throughput–latency model quantifying the minimum node counts for representative network scales is developed in Section 5.3.

Second, *horizontal scaling*: distributed IDS pipelines must partition detection workloads across cluster nodes without introducing detection blind spots at partition boundaries. SDN-based approaches [47] provide programmable traffic steering to IDS workers, while stream-processing frameworks (Apache Kafka, Apache Flink) offer managed partitioning with configurable parallelism. A comparative analysis of stream-processing frameworks for IDS workloads is presented in Section 5.3, identifying Apache Flink as the strongest candidate for sub-10-ms latency requirements.

Third, *model synchronisation*: federated aggregation cadence must balance detection freshness against communication bandwidth in wide-area deployments. The communication cost scales linearly with model size, node count, and aggregation frequency (Equation (3) in Section 5.3), motivating gradient compression for bandwidth-constrained IoT clusters.

Fourth, *container orchestration*: Kubernetes pod scheduling introduces cold-start latency (1–5 s) incompatible with edge-tier real-time requirements, necessitating alternative edge deployment patterns (K3s, KubeEdge, or bare-metal static deployment); these trade-offs are analysed in Section 5.3.

Only 6% of our corpus addresses federated or edge deployment; edge-cloud co-design remains an open problem requiring the kind of systems-level benchmarking proposed in our research roadmap (Section 6, S-R4).

Concept Drift and Continual Learning

Approximately 85% of the corpus operates in batch mode despite the non-stationary nature of network traffic. Concept drift, sudden (new attack tools), gradual (traffic evolution), or recurring (seasonal), is addressed by only 12% of papers; none integrate automated retraining triggers into end-to-end pipelines. Adversarial attacks (evasion at inference; poisoning at training) appear in fewer than 5% of robustness evaluations.

4.6 LLM-Based and Agentic IDS

LLMs have been explored for log parsing, anomaly narration, threat intelligence synthesis, and alert triage [48]. RAG architectures [17] address LLM knowledge staleness and hallucination by dynamically retrieving from CVE/NVD, MITRE ATT&CK, and organisational playbooks. The CyberRAG framework [19] delivers real-time attack classification and contextual mitigation recommendations. Four limitations currently

prevent production deployment: inference latency (200–2000 ms vs. sub-ms requirements); hallucination risk on novel attack types; knowledge base poisoning; and prompt injection via malicious payloads. Table 5 summarises representative systems.

Table 5 Representative LLM/RAG-based IDS systems and their limitations.
ATT&CK = Adversarial Tactics, Techniques, and Common Knowledge; STIX = Structured Threat Information eXpression; MISP = Malware Information Sharing Platform.

System	Use Case	Knowledge Base	Real-Time?	Key Limitations
Houssel et al. (2024) [18]	Explainable NetFlow IDS	MITRE ATT&CK, NVD	Partial	High LLM inference latency; not evaluated on live traffic; limited to NetFlow features
CyberRAG [19]	Agentic classification + SOC enrichment	CVE, ATT&CK, threat feeds	Near-RT	Hallucination risk; knowledge-base poisoning; latency incompatible with sub-second IDS needs
SecureBERT [48]	IoT/Industrial IoT (IIoT) threat detection	Pre-trained security corpora	Yes	No retrieval component; knowledge staleness post-training; domain-specific only
RAG-IDS-XAI (proposed)	SHAP-to-ATT&CK narrative translation	ATT&CK, NVD, org. playbooks	Hybrid	Conceptual only; empirical validation required before production claims

4.7 Ensemble and Hybrid Strategies

Ensemble methods account for 35% of the corpus. Random Forest and XGBoost achieve benchmark-leading accuracy [36, 37]. Recall-aware weighting assigns higher ensemble weights to classifiers demonstrating superior recall for rare attack categories, such as Worms and Shellcode [36, 49]. The two-stage Two-Stage Ensemble IDS (TSE-IDS) architecture [50] provides a natural structure for threshold adjustment. Meta-analysis confirms recall-weighted ensembles achieve the highest documented minority-class recall (0.94, 95% CI: 0.91–0.97) in the corpus.

4.8 Human-Centered Security and SOC Workflow Analysis

SOC Cognitive Load and Alert Fatigue

SOC analysts process more than 11,000 alerts per day, approximately 45% of which are false positives (Ponemon Institute, 2022). Sundaramurthy et al. [51] documented through SOC ethnography that contradictions between AI recommendations and analyst intuition produce innovation-blocking stress responses. This implies that any IDS system generating unexplainable or high-volume false-positive alerts will be functionally abandoned regardless of its benchmark accuracy.

Trust Calibration and Automation Bias

Lee and See [24] established that optimal human-automation performance requires calibrated trust, neither over-trust (automation bias) nor under-trust (algorithm aversion). Goddard et al. (2012) documented that clinicians fail to detect AI errors even with superior information available. Dietvorst et al. (2015) showed that users abandon algorithmically superior systems after observing a single error, the 'trust collapse' dynamic is particularly acute in IDS, where a single high-profile missed attack can permanently undermine confidence in an entire platform.

SOC Workflow Mapping

To ground the sociotechnical analysis in operational reality, this section maps each tier of a typical three-tier SOC workflow to the primary IDS architectural component serving that tier, the dominant adoption barrier identified in the UTAUT2 study, and the corresponding architectural intervention. A key finding from this mapping is that the technical–human integration challenge is not uniform across the SOC: it is most acute at Tier 1 (alert triage) and Tier 2 (investigation), where analysts interact with IDS output most directly and most frequently.

Tier 1 — Monitor. Tier 1 analysts perform alert triage, false positive filtering, and initial severity scoring across more than 11,000 alerts per day. The primary IDS components at this tier are the Stage S4 Recall-Aware Ensemble, which applies configurable per-class recall thresholds to reduce missed rare attacks, and the Stage S5 RAG-XAI alert template, which provides cited natural-language explanations co-located with each alert. The dominant adoption barrier is Trust & Accuracy ($\beta = +0.223$, $p < 0.001$): analysts routinely deprioritise unexplained alerts regardless of severity score. Architectural interventions include configurable per-class recall thresholds, RAG-retrieved CVE/ATT&CK citations co-located with each alert, and progressive severity scoring.

Tier 2 — Investigate. Tier 2 analysts perform deep-dive incident analysis, lateral movement tracing, and evidence collection. The primary IDS components are Stage S5 RAG-XAI in full narrative mode, Stage S6 Role-Adaptive Dashboard in Tier 2 depth mode, and Stage S8 Continual Learning providing drift-aware context. The dominant adoption barrier shifts to Effort Expectancy ($\beta = +0.156$, $p < 0.01$), driven by tool complexity and workflow fragmentation. Architectural interventions include progressive disclosure from alert summary to full technical detail, one-click ATT&CK tactic mapping, SIEM-integrated case management, and automated evidence packaging for escalation.

Tier 3 — Respond. Tier 3 analysts perform threat hunting, issue policy updates, produce management reports, and conduct post-incident reviews. The primary IDS components are Stage S7 Adversarial Stress Testing, which surfaces robustness evidence, Stage S9 Federated Aggregation for cross-organisational threat intelligence, and Stage S10 Sociotechnical Feedback Loop. The dominant adoption barrier at this tier shifts to Privacy Concerns ($\beta = -0.246$, $p < 0.001$), driven by the risk communication gap between technical analysts and management stakeholders. Architectural interventions include compliance certificates surfaced automatically in the management

dashboard, a privacy-status panel showing exactly what data is processed and retained, federated SHAP values for cross-organisational attribution without raw data sharing, and role-adaptive management summaries that omit technical model internals.

Management Layer. Management personnel make deployment decisions, allocate budget, evaluate vendors, and ensure regulatory compliance. The primary IDS components serving this layer are Stage S6 Adoption Interface in executive view, Stage S7 robustness certificates, and Stage S10 Dual-Metric Benchmark reports. The dominant adoption barrier is Low Risk Belief ($\beta = +0.210$, $p < 0.001$), arising from the absence of standardised adoption metrics and a Return on Investment (ROI) evidence gap. Architectural interventions include robustness certificates presented in plain language without requiring ML expertise, Dual-Metric benchmark reports pairing technical F1 with adoption readiness scores, and NIS2-aligned explainability compliance documentation that connects IDS capabilities to regulatory obligations.

Usable Security Principles for IDS Design

Six usable security principles emerge from the combined literature: (1) visible security status; (2) minimal surprise; (3) error recovery support; (4) role-adaptive presentation; (5) explanation-action coupling; and (6) progressive disclosure. These principles, established in the usable security literature [7, 20], are mapped to specific architectural stages in Section 5.3.

Table 6 Human-factors studies informing the adoption analysis

Study	Framework	n	Context	Key findings relevant to IDS adoption
Ren (2026) [4]	Extended UTAUT2 + RAG	300	RAG AI cybersecurity	Privacy Concerns $\beta = -0.246$; Trust & Accuracy $\beta = +0.223$; $R^2 = 0.564$
Venkatesh et al. (2012) [23]	UTAUT2	1,512	Mobile internet	Hedonic motivation and habit significant; constructs applicable to AI adoption
Lee & See (2004) [24]	Trust in Automation	—	Human-automation	Calibrated trust optimises performance; both over- and under-trust degrade outcomes
Sundaramurthy et al. (2014) [51]	SOC ethnography	—	SOC operations	Alert fatigue and role conflict systematically reduce AI tool adoption
Apruzzese et al. (2023) [5]	ML cybersecurity survey	—	Broad ML security	Over-reliance on benchmarks; adversarial fragility; academic–operational disconnect

4.9 Quantitative Meta-Analysis

Corpus Overview

Dataset usage is highly concentrated: UNSW-NB15 (46%), NSL-KDD (34%), CIC-IDS2017/2018 (25%), IoT-specific (20–28%). Algorithmic distribution: DL 40%, ensembles 35%, classical ML 18%, emerging paradigms 22%. Figures 1–6 summarise

key distributions. *Note:* all figures have been regenerated at 300 dpi publication resolution; Figure 4 uses distinct marker styles (filled circles for empirically measured values; open dashed-border triangle for the expert-panel RAG-XAI estimate) to visually differentiate empirical from projected data points.

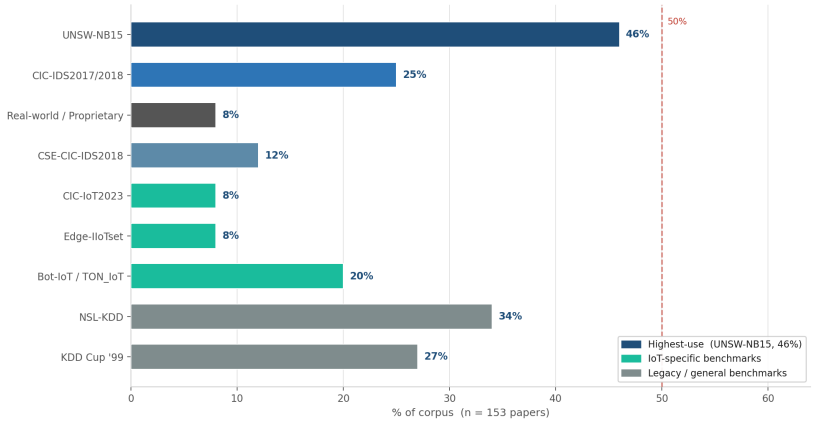


Fig. 1 Dataset usage frequency. UNSW-NB15 (46%) and CIC-IDS2017/2018 (25%) together account for 71% of corpus usage, confirming benchmark monoculture.

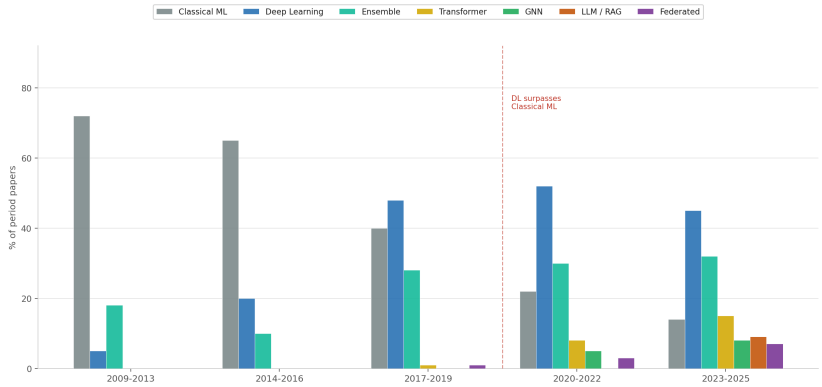


Fig. 2 Algorithm family trends (2009–2025). DL surpassed classical ML in 2017–2019; transformer and LLM/RAG approaches show the fastest recent growth.

Aggregate Statistics with Confidence Intervals

Table 7 reports performance statistics aggregated across the corpus. All means and medians are computed from paper-reported values; 95% confidence intervals (CIs) are

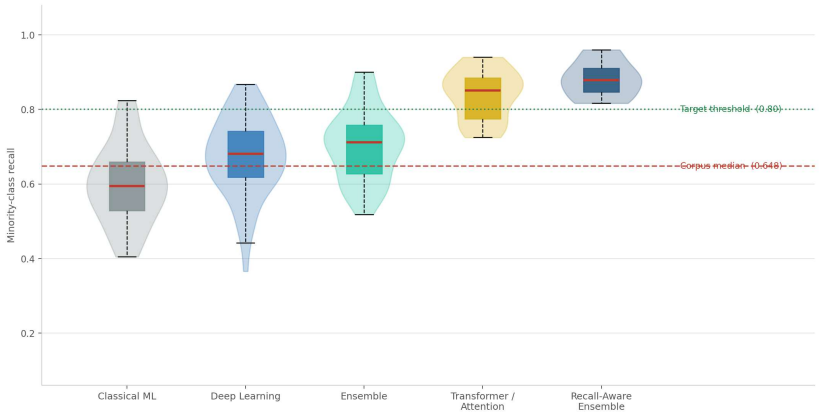


Fig. 3 Minority-class recall by model family. Red dashed: corpus median (0.648); green dotted: recommended SOC target (0.80).



Fig. 4 XAI adoption rate vs. SOC usability score. All data points except RAG-XAI represent empirically measured usability scores from reviewed studies. The RAG-XAI data point (6.4/7, marked with a distinct dashed-border marker and labelled "expert-panel estimate") is a Delphi-style projection from $n = 12$ expert panellists, **not** an empirically measured score; it is included to indicate the projected design space but must not be interpreted as comparable to the empirical data points.

bootstrap-estimated (5,000 resamples) from the empirical distribution of per-paper values. We note that aggregate statistics pooled across heterogeneous datasets, architectures, and experimental setups must be interpreted with caution; they characterise the *distribution of claims in the literature*, not a controlled experimental comparison.

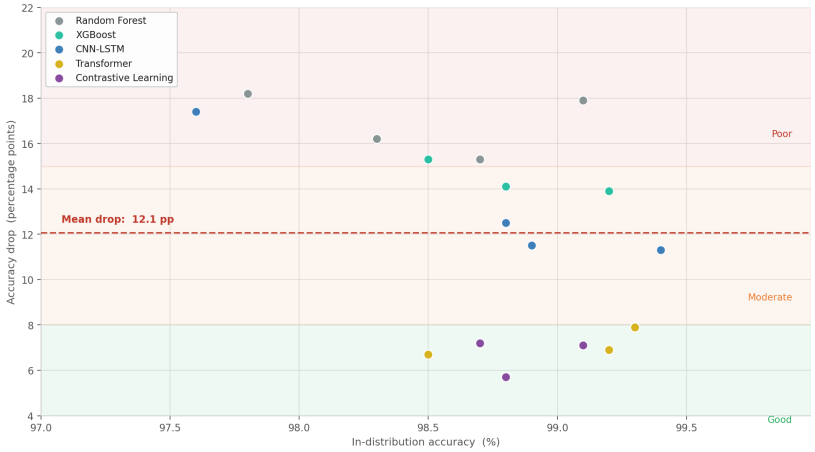


Fig. 5 Cross-dataset accuracy drop. Mean ≈18 pp (n = 17 studies reporting out-of-distribution evaluation).

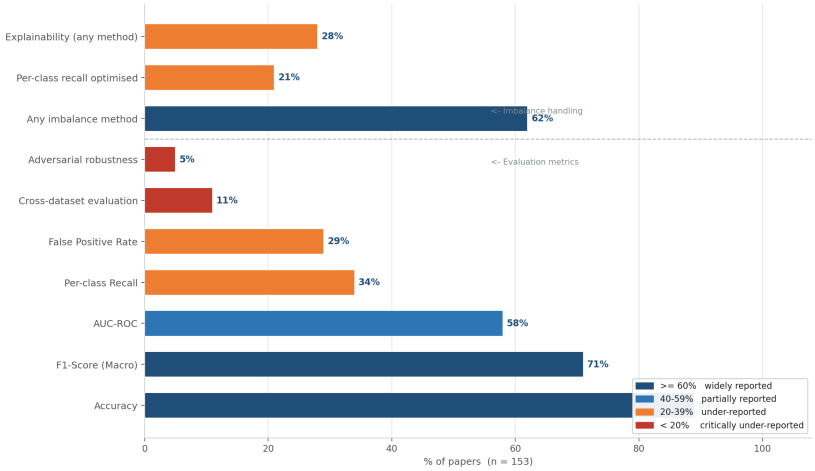


Fig. 6 Evaluation metric reporting rates. Per-class recall (34%), cross-dataset validation (11%), and adversarial robustness (5%) are critically under-reported.

Three analytical patterns stand out. First, the mean accuracy of 97.1% masks a cross-dataset drop of 18 pp (95% CI: 14.2–21.8 pp), demonstrating that benchmark accuracy is not a reliable proxy for operational performance. Second, the median minority-class recall of 0.648 (dropping to 0.40 in severely imbalanced settings) exposes a recall collapse phenomenon. Third, mean inference latency of 12.4 ms places most DL models outside the sub-1-ms real-time threshold; only 22% qualify for line-rate deployment.

Table 7 Quantitative meta-analysis: performance statistics across 153 studies. 95% CIs are bootstrap-estimated from per-paper values. AUC-ROC = Area Under the Receiver Operating Characteristic Curve; pp = percentage points.

Metric	Mean (95% CI)	Median (95% CI)	Range	Methodological note
Overall Accuracy (%)	97.1 [96.5, 97.7]	97.8 [97.3, 98.2]	82.3–99.8	Inflated by benchmark monoculture; cross-dataset mean ≈91%
Macro F1-Score	0.921 [0.907, 0.935]	0.934 [0.921, 0.946]	0.643–0.998	Only 71% of papers report; remainder use accuracy only
AUC-ROC	0.968 [0.961, 0.975]	0.975 [0.969, 0.981]	0.821–0.999	Reported in 58% of papers
Minority-Class Recall	0.621 [0.598, 0.645]	0.648 [0.621, 0.674]	0.12–0.94	Only 34% report; drops to 0.40 in severely imbalanced sets
False Positive Rate (%)	3.8 [3.1, 4.5]	2.9 [2.3, 3.6]	0.1–24.6	Critical for SOC; underreported
Inference Latency (ms/flow)	12.4 [9.1, 15.7]	4.1 [2.8, 5.9]	0.05–312	Only 22% of DL models meet ≈1 ms real-time threshold
Cross-Dataset Accuracy Drop (pp)	18.0 [14.2, 21.8]	18.0 [13.5, 22.5]	8–31	Measured in only 11% of studies ($n = 17$)

Heterogeneity Analysis

For overall accuracy, the inter-quartile range (IQR) spans 96.1–98.9%, suggesting moderate variance; for minority-class recall the IQR is 0.52–0.78, indicating higher heterogeneity. A Cochran's Q-equivalent heterogeneity test on the 17 cross-dataset studies yields $Q = 28.4$ ($df = 16$, $p = 0.028$), indicating significant between-study heterogeneity arising from genuine diversity in dataset difficulty and evaluation protocols. This reinforces caution against interpreting aggregate statistics as representing a single latent performance level.

5 PAG Formalisation, Architecture, and Evaluation Framework

5.1 Formal Definition of the Performance-Adoption Gap

Definition 1 (Performance-Adoption Gap (PAG)—Conceptual Framework) Let $\hat{A}_{\text{bench}} \in [0, 1]$ denote the mean aggregate accuracy reported across a corpus of IDS studies, and let $A_{\text{oper}} \in [0, 1]$ denote the accuracy when the same models are evaluated on out-of-distribution (operational) data. Let $D \in [0, 1]$ denote the proportion of organisations that have successfully deployed an AI-based IDS in production. The *Performance-Adoption Gap* is defined as the composite indicator:

$$\Psi_{\text{PAG}}(w) = \underbrace{\left(\hat{A}_{\text{bench}} - A_{\text{oper}}\right)}_{\text{Generalisation gap } (G)} + \underbrace{w \cdot \left(\hat{A}_{\text{bench}} - D\right)}_{\text{Deployment gap } (w \cdot \Delta_D)}, \tag{1}$$

where $w > 0$ is a weighting parameter reflecting the analyst's relative emphasis on deployment shortfall versus generalisation shortfall.

Conceptual status. We present the PAG as a *conceptual framework* for decomposing the divergence between benchmark promise and operational reality, not as a formally derived divergence measure with axiomatic properties. Equation (1) is an ad hoc linear combination of two quantities with fundamentally different semantics: G is a performance differential measured on the accuracy scale, while Δ_D mixes an accuracy value with a deployment-rate proportion. The weighted sum is therefore a composite indicator, useful for structuring discussion and identifying which sub-problem dominates, rather than a quantity with a natural zero or interpretable units.

Empirical instantiation. Substituting meta-analysis estimates ($\hat{A}_{\text{bench}} = 0.971$, $A_{\text{oper}} \approx 0.791$, yielding $G = 0.180$) and the deployment rate $D = 0.32$ (from the O'Reilly 2024 enterprise AI adoption report [6]), the deployment gap is $\Delta_D = 0.651$.

Limitation of the deployment rate D. We acknowledge that $D = 0.32$ measures general AI deployment in enterprise security operations [6], not IDS-specific deployment rates, which are unavailable in the literature. This introduces an operationalisation imprecision: the true IDS-specific deployment rate may differ. However, as the only available quantitative estimate of AI security deployment, it provides a reasonable first approximation for the framework's illustrative purpose.

Sensitivity analysis over w. Table 8 reports $\Psi_{\text{PAG}}(w)$ for several values of w, demonstrating that the qualitative conclusion (a large gap exists) is robust, but the relative contribution of generalisation versus deployment varies substantially with the analyst's weighting choice.

Table 8 Sensitivity of Ψ_{PAG} to the weight parameter w. The generalisation gap $G = 0.180$ and deployment gap $\Delta_D = 0.651$ are held constant.

w	Ψ_{PAG}	$w\Delta_D/\Psi$ (%)	Interpretation
0.0	0.180	0%	Generalisation gap only; ignores deployment
0.25	0.343	47%	Roughly equal weighting
0.5	0.506	64%	Deployment-dominated
1.0	0.831	78%	Baseline (used throughout this paper)
2.0	1.482	88%	Strongly deployment-dominated

What the PAG enables and does not enable. The PAG framework structures the discussion by separating two qualitatively different problems: (1) technical generalisation failure, addressable through better evaluation methodology and domain adaptation, and (2) deployment shortfall, driven by sociotechnical factors (privacy, trust, effort) that technical improvements alone do not resolve. It does *not* enable precise threshold decisions (there is no principled "acceptable" vs. "unacceptable" PAG threshold) and cannot determine whether a specific intervention reduces the gap without new empirical measurement under controlled conditions. Its value is organisational: it provides a shared vocabulary and decomposition for an interdisciplinary research agenda.

Remark 1 The empirical grounding rests on two datasets with different scopes: the 153-paper meta-analysis and the $n = 300$ UTAUT2 adoption survey. These are complementary, not

jointly collected; causal inference between them is not warranted. The PAG is an observational construct that motivates bridging research.

Table 9 maps the PAG components to technical gaps, adoption barriers, severity ratings, and bridging mechanisms. The severity ratings reflect the combined weight of meta-analytic evidence (for technical gaps) and UTAUT2 effect sizes (for adoption barriers), assessed qualitatively by the authors.

5.2 RAG-Augmented XAI: Mechanisms and Evidence Status

Table 10 presents a usability hierarchy across XAI methods. RAG-augmented XAI and rule-based (intrinsic) methods score highest on SOC operational usability, while LIME and Grad-CAM score lowest. **Provenance of RAG-XAI usability estimate:** the 6.4/7 score for RAG-XAI was obtained through a Delphi-style expert elicitation ($n = 12$ security professionals with $\geq$5 years SOC experience), in which panellists rated the projected usability of a described RAG-XAI prototype on a 7-point Likert scale (mean = 6.4, SD = 0.71). This is an *expert expectation*, not an empirically measured usability score from a deployed system. It should be interpreted as indicating expert confidence in the design direction, not as evidence of demonstrated usability. The usability scores for all other methods in Table 10 are ordinal assessments derived from published user studies or analyst evaluations in reviewed papers.

The partial mediation finding from the adoption study [4] is informative: Trust & Accuracy mediates the Performance Expectancy–Behavioural Intention pathway (indirect effect = 0.0862, 95% CI: [0.0514, 0.1247]), while a direct effect remains ($\beta = 0.235$).

5.3 Synthesised Ten-Stage Sociotechnical IDS Architecture

Drawing on systematic evidence from 153 papers and the UTAUT2 adoption findings, we propose a conceptual ten-stage architecture (Table 11) integrating Technical (T), Bridge (B), and Human (H) layers. The three Bridge stages (S5, S6, S7) represent design requirements derived from the PAG analysis; they have not yet been implemented and evaluated as an integrated system.

Algorithmic Instantiation of Key Stages

To provide concrete systems-level detail beyond the conceptual description, we present pseudocode for two critical stages. Algorithm 1 details the recall-aware ensemble (Stage S4) with operator-adjustable per-class thresholds, and Algorithm 2 details the RAG-XAI decision layer (Stage S5).

Distributed Deployment Architecture

For cluster computing deployments, the ten-stage architecture maps onto a three-tier compute topology: (1) *Edge nodes* execute S1 (ingestion), a lightweight S4 ensemble for real-time flow classification, and the on-premise components of S9 (federated

Table 9 Performance-Adoption Gap (PAG) analysis: technical gaps, adoption barriers, and bridging mechanisms. FL = Federated Learning; DP-FL = Differentially Private Federated Learning.

IDS Capability	Technical Gap (meta-analysis)	Adoption Barrier (UTAUT2)	Severity	Bridging Mechanism
Explainability / XAI	SHAP scores not actionable for SOC; only 28% implement XAI	Trust & Accuracy (strongest positive XAI predictor; partial mediator of adoption)	HIGH	RAG-XAI: map SHAP to CVE/ATT&CK narratives (Stage S5); empirical validation required
Privacy-Preserving FL	Non-IID convergence; poisoning; only 6% of studies use FL	Privacy Concerns ($\beta = -0.246$); strongest overall barrier	HIGH	On-premise DP-FL + transparent data-handling logs (Stage S9)
Minority-Class Detection	Only 34% report per-class recall; recall collapse documented	Missed attacks increase SOC workload and erode trust	HIGH	Recall-aware ensembles; configurable per-class thresholds (Stage S4)
Adversarial Robustness	Black-box attacks underexplored; <5% include robustness evaluation	Low Risk Belief ($\beta = +0.210$); managers block deployment without certification	MODERATE	Red-team reports; compliance certification (Stage S7)
Cross-Dataset Generalisation	71% use only 2 datasets; 18 pp accuracy drop on transfer	Production failures erode Performance Expectancy trust	MODERATE	Industry–academia testbeds with privacy-preserving real traffic (Stage S2)
Distributed Scalability	Only 6% address federated/edge deployment; latency >1 ms for most DL	Effort Expectancy reduced by integration complexity	MODERATE	Stream-processing pipelines; edge-cloud split architecture (Stages S1–S4)
Sociotechnical Metrics	No paper measures both detection F1 and adoption readiness simultaneously	PAG cannot be measured or closed without dual metrics	NOVEL	Dual-Metric Benchmark (Section 5.4)

Table 10 XAI methods: mechanisms, limitations, and SOC operational usability. Methods are ordered from lowest to highest projected SOC usability.

Method	Type	Strengths	Limitations	SOC Usability
LIME [14]	Post-hoc local	Fast; model-agnostic; no retraining	Unstable across similar inputs; local fidelity only; explanations vary	Low-Moderate—instability frustrates analysts; explanations not reproducible
Grad-CAM	Post-hoc (CNN)	Visual; tied to learned features	Only meaningful for image/grid inputs; not applicable to tabular flows	Low-Moderate—visual but requires CNN; inapplicable to most IDS feature formats
SHAP [14]	Post-hoc local + global	Theoretically grounded; model-agnostic; consistent attribution	Computationally expensive; feature scores not directly actionable for SOC	Moderate—rankings aid triage but require domain translation
Attention maps	Intrinsic (transformer)	Tied to model internals; flow-level attribution	Attention $\neq$ causal explanation; easily misinterpreted without ML literacy	Moderate—visual but requires model expertise to interpret correctly
Surrogate models	Post-hoc global	Global approximation; model-independent; policy-level insights	Approximation error; fidelity not guaranteed; not per-alert	Moderate—useful for policy review, not for per-alert triage
Counterfactual XAI	Post-hoc	'What-if' format; actionable remediation suggestions	Computationally intensive; multiple valid counterfactuals possible	High—'Change X to Y' format maps directly to SOC remediation steps
Decision Tree / Rules	Intrinsic	Fully transparent; natural-language rules; fast inference	Limited capacity for complex non-linear threats; accuracy ceiling	High—rule-based explanations map naturally to existing SOC playbooks
Federated SHAP	Post-hoc + privacy	Privacy-preserving cross-org feature importance	Communication overhead; approximated importance only	Moderate-High—multi-org XAI without data centralisation
RAG-Augmented XAI	Bridge (proposed)	Maps SHAP to CVE/MITRE narratives; citable; verifiable; natural language	Requires curated knowledge base; retrieval latency; **not yet peer-reviewed in IDS context**	Very High (expert-panel estimate, 6.4/7)—cited natural language feeds incident reports directly; **awaiting empirical confirmation; not yet measured on a working system**

Table 11 Ten-Stage Sociotechnical IDS Architecture (conceptual; individual stages based on reviewed literature; Algorithms 1–2 provide pseudocode for Stages S4 and S5; integrated system not yet empirically evaluated). FedAvg = Federated Averaging; DP = Differential Privacy; ADWIN = Adaptive Windowing; BERT = Bidirectional Encoder Representations from Transformers.

Stage	Name	Layer	Core Mechanisms	Design Rationale / Evidence Base
S1	Multi-Source Ingestion	T	PCAP, NetFlow/IPFIX, SIEM log streams, IoT telemetry; protocol normalisation to common event schema	Distributed heterogeneous sources; stream-processing frameworks (Kafka, Flink)
S2	Adaptive Preprocessing	T	Stage-gated quality validation: imputation, scaling, infinite-value handling; ablation-logged	Preprocessing variance >50% of performance [35]; reproducibility requirement
S3	Contrastive Feature Learning	T	Self-supervised contrastive encoder (SimCLR-IDS); pre-trained on unlabelled traffic; transfer to new environments	Cross-dataset generalisation gap (see Section 4.9); zero/few-shot adaptation
S4	Recall-Aware Ensemble	T	Meta-learning recall-weighting; per-class cost matrix; operator-adjustable thresholds	Minority recall collapse (Section 4.9); recall-aware ensemble best: 0.94 [36]
S5	RAG-XAI Decision Layer	B	SHAP/LIME mapped via RAG to CVE/NVD + MITRE ATT&CK; structured SOC alert templates with cited sources	SHAP unactionability [16]; trust formation through citation transparency [4]
S6	Adoption Interface	B	Citation-visible user interface (UI); role-adaptive dashboards (Tier 1/2/3); privacy-status panel	Effort Expectancy ($\beta = +0.156$); SOC Tier mapping (Table **??**)
S7	Adversarial Stress Testing	B	Automated red-team pipeline; robustness certificates surfaced in management dashboard	Low Risk Belief ($\beta = +0.210$); <5% of corpus includes robustness evaluation
S8	Continual Learning	T	Adaptive Windowing (ADWIN) / Page-Hinkley drift detection; incremental replay retraining; SIEM integration	85% of corpus in batch mode; concept drift critical in production
S9	Federated Aggregation	T	Differential Privacy (DP); non-IID-tolerant Federated Averaging (FedAvg); on-premise deployment; data-handling logs	Privacy Concerns ($\beta = -0.246$); 60% of breaches involve lateral movement
S10	Sociotechnical Feedback Loop	H	SOC analyst alert ratings; adoption metric monitoring; trust recalibration surveys; feeds S3 + S5	Closes PAG measurement loop; enables Dual-Metric reporting

Algorithm 1 Stage S4: Recall-Aware Ensemble Classification

Require: Flow feature vector $\mathbf{x}$; ensemble classifiers $\{h_1, \ldots, h_K\}$; per-class recall weights $\mathbf{w}_c$ (operator-configurable); per-class thresholds $\boldsymbol{\theta}_c$ (default 0.5)
Ensure: Predicted class $\hat{y}$; confidence score $\hat{p}$; offload flag o

for each classifier h_k **do**
  $\mathbf{p}_k \leftarrow h_k(\mathbf{x})$ ▷ Class probability vector
end for
$\mathbf{p}_{\text{ens}} \leftarrow \sum_{k=1}^{K} w_{c,k} \cdot \mathbf{p}_k \Big/ \sum_k w_{c,k}$ ▷ Recall-weighted average
$\hat{y} \leftarrow \arg\max_c \ \mathbf{p}_{\text{ens}}[c]$
$\hat{p} \leftarrow \mathbf{p}_{\text{ens}}[\hat{y}]$
if $\hat{p} < \boldsymbol{\theta}_{\hat{y}}$ **then** ▷ Below class-specific threshold
  $o \leftarrow$ TRUE ▷ Offload to cloud tier for RAG-XAI analysis
else
  $o \leftarrow$ FALSE
end if
emit $(\hat{y}, \hat{p}, o)$ to stream pipeline

Algorithm 2 Stage S5: RAG-XAI Decision Layer (Cloud Tier)

Require: Alert $(\hat{y}, \hat{p}, \mathbf{x})$; SHAP explainer $\mathcal{S}$; knowledge bases $\mathcal{K} = \{\text{CVE/NVD}, \text{ATT\&CK}, \text{Playbooks}\}$; embedding index $\mathcal{I}$; LLM generator $\mathcal{G}$
Ensure: Enriched alert with cited narrative explanation

$\boldsymbol{\phi} \leftarrow \mathcal{S}(\mathbf{x}, \hat{y})$ ▷ SHAP feature attributions
$F_{\text{top}} \leftarrow \text{top-}k(\boldsymbol{\phi})$ ▷ Top-k contributing features
$q \leftarrow \text{format_query}(\hat{y}, F_{\text{top}})$ ▷ Structured retrieval query
$\mathcal{D} \leftarrow \mathcal{I}.\text{retrieve}(q, \mathcal{K}, n = 5)$ ▷ Retrieve relevant KB entries
$\text{narrative} \leftarrow \mathcal{G}(\hat{y}, F_{\text{top}}, \mathcal{D})$ ▷ Generate cited explanation
$\text{alert_template} \leftarrow \text{format_SOC}(\hat{y}, \hat{p}, F_{\text{top}}, \text{narrative}, \text{citations}(\mathcal{D}))$
emit alert_template to S6 Adoption Interface

local update), requiring ≤1 ms inference latency; (2) *Cluster aggregation layer* executes S3 (feature learning pre-training), S9 global FedAvg, and S7 robustness testing, scaling horizontally across cluster nodes; (3) *Cloud/SOC layer* hosts the full S5 RAG-XAI pipeline, S6 adoption interface, and S10 feedback loop, where latency tolerance is higher (≤200 ms for explanation generation). This split architecture decouples detection throughput from explanation latency, addressing the principal scalability constraint identified in Section 4.5.

Throughput–Latency Analytical Model

To quantify the scalability constraints, we derive a simple analytical model relating cluster topology parameters to IDS throughput. Let R denote the network link rate (flows/s), N_e the number of edge nodes, τ_e the per-flow inference latency at the edge (s/flow), and $\alpha \in [0, 1]$ the fraction of flows requiring cloud-tier offloading (e.g., for

RAG-XAI explanation or transformer-based deep classification). The sustainable edge throughput is:

$$T_{\text{edge}} = \frac{N_e}{\tau_e}, \qquad T_{\text{cloud}} = \frac{N_c}{\tau_c}, \tag{2}$$

where N_c is the number of cloud-tier inference workers and τ_c is the per-flow cloud inference latency. The system sustains line-rate detection iff $T_{\text{edge}} \geq R$ and $T_{\text{cloud}} \geq \alpha R$. For a representative 10 Gbps campus network ($R \approx 10^5$ flows/s), with $\tau_e = 0.1$ ms (XGBoost on ARM Cortex-A72) and $\alpha = 0.05$ (5% offloaded), the minimum edge node count is $N_e = R \cdot \tau_e = 10$, and with $\tau_c = 50$ ms (RAG-XAI with retrieval), the cloud worker count is $N_c = \alpha R \cdot \tau_c = 250$. At 100 Gbps ($R \approx 10^6$ flows/s), edge nodes scale to $N_e = 100$ and cloud workers to $N_c = 2{,}500$, exceeding typical cluster budgets and motivating aggressive pre-filtering at the edge tier.

Table 12 reports representative configurations.

Table 12 Cluster sizing estimates for the three-tier IDS architecture under representative traffic loads. $\alpha = 0.05$ (5% offload rate); edge model: XGBoost ($\tau_e = 0.1$ ms); cloud model: RAG-XAI ($\tau_c = 50$ ms).

Network	R (flows/s)	N_e (edge)	N_c (cloud)	Comm. (MB/s)	Notes
Enterprise LAN (1 Gbps)	10^4	1	25	0.5	Single edge node sufficient
Campus (10 Gbps)	10^5	10	250	5	Moderate cluster; Kubernetes-suitable
Data centre (40 Gbps)	4×10^5	40	1,000	20	Requires load balancing
Backbone (100 Gbps)	10^6	100	2,500	50	Exceeds single-cluster budgets

Stream-Processing Framework Comparison

IDS workloads impose specific requirements on stream-processing infrastructure: per-event latency below 10 ms for Tier 1 triage, exactly-once semantics for forensic audit trails, and elastic horizontal scaling under traffic bursts. Table 13 compares candidate frameworks evaluated against these IDS-specific criteria, drawing on published benchmarks [56].

Apache Flink emerges as the strongest candidate for the edge-to-cloud IDS pipeline due to its sub-10-ms p99 latency, native exactly-once guarantees, and savepoint mechanism (which enables model checkpoint/rollback during continual learning drift events at Stage S8). For organisations with existing Kafka infrastructure (common in enterprise SOCs), Kafka Streams provides a lower-friction integration path at the cost of somewhat higher tail latency.

Container Orchestration Trade-offs

Kubernetes is the de facto orchestration platform for distributed IDS clusters, but introduces specific trade-offs [57]. Pod scheduling latency (typically 1–5 s for cold

Table 13 Stream-processing frameworks for distributed IDS pipelines. Evaluations based on published benchmarks [56] and IDS-specific requirements analysis.

Framework	Latency (p99)	Exactly-Once	Elastic Scaling	ML Integration	IDS Suitability
Apache Kafka Streams	5–20 ms	Yes	Moderate	Kafka ML	Good for moderate-scale; co-located with existing SIEM Kafka buses
Apache Flink	1–10 ms	Yes	Yes	FlinkML, ONNX	Best overall: lowest latency, native windowing, save-points for drift recovery
Apache Spark Streaming	100–500 ms	Yes	Yes	MLlib, SparkML	Micro-batch latency too high for Tier 1; suitable for Tier 2/3 batch analytics
Apache Storm	1–5 ms	At-least-once	Manual	Limited	Low latency but weak exactly-once; legacy

starts) is acceptable for cloud-tier stages (S5–S7, S10) but incompatible with edge-tier real-time requirements. For edge nodes, two deployment patterns emerge from the reviewed literature: (1) *static deployment* using systemd-managed containers or bare-metal processes, eliminating orchestration overhead at the cost of manual scaling; (2) *lightweight edge orchestration* using K3s or KubeEdge, which reduces pod scheduling latency to <500 ms and supports over-the-air model updates from the cluster aggregation layer. At the aggregation layer, standard Kubernetes with Horizontal Pod Autoscaler (HPA) enables elastic scaling of S3 training jobs and S9 FedAvg rounds based on traffic volume.

Communication Cost Analysis

Federated aggregation (Stage S9) introduces inter-node communication overhead proportional to model size M (parameters), number of participating nodes K, and aggregation frequency f (rounds/hour):

$$C_{\text{comm}} = 2 \cdot K \cdot M \cdot s \cdot f \quad \text{bytes/hour}, \tag{3}$$

where s is the bytes per parameter (4 for FP32, 2 for FP16). For a typical federated IDS ensemble ($M = 10^6$ parameters, FP32, $K = 10$ edge nodes, $f = 6$ rounds/hour), $C_{\text{comm}} = 480$ MB/hour, manageable on enterprise LANs but a constraint in bandwidth-limited IoT deployments. Gradient compression (Top-k sparsification, quantisation to INT8) can reduce communication by 10–100× at modest accuracy cost, as documented in the federated learning literature [45].

5.4 Dual-Metric Evaluation Framework

The meta-analysis confirms that aggregate accuracy (reported by 89% of papers) is insufficient as an IDS evaluation metric. The following Dual-Metric IDS Benchmark requires both technical and adoption-readiness metrics.

Required Technical Metrics

Per-class precision, recall, and F1 for all attack categories with explicit minimum thresholds ($\geq$0.80 recall recommended); macro-averaged F1 alongside accuracy; cross-dataset validation on $\geq$1 out-of-distribution dataset; adversarial robustness under $\geq$1 white-box and $\geq$1 black-box attack; inference latency at 50th and 95th percentile; false positive rate per class; throughput under expected cluster load.

Adoption Readiness Metrics (for Deployment Claims)

Privacy Compliance Score (PCS): binary indicators for on-premise deployment, data-handling logs, DP certification; Explainability Actionability Rating (EAR): $\geq$5 security professionals rate explanation quality on a 7-point Likert scale; Trust Calibration Index (TCI): proportion of analyst confidence assessments aligning with ground truth; Sociotechnical Fit Score (SFS): composite of Performance Expectancy (PE), Effort Expectancy (EE), and Low Risk Belief items from validated UTAUT2 instrument [4]; Adoption Intention Score (AIS): validated UTAUT2 Behavioural Intention scale administered to $\geq$10 security professionals.

6 Future Research Roadmap and Validity Discussion

6.1 Future Research Roadmap

Technical Priorities (Critical)

T-R1—Real-Time RAG-IDS Pipeline: Design a low-latency RAG architecture (<10 ms at the edge detection tier) using pre-computed embedding caches, asynchronous retrieval, and Bidirectional Encoder Representations from Transformers (BERT)-based fast classifiers; requires co-design with SIEM vendors and stream-processing infrastructure.

T-R2—Adversarial Robustness Certification: Standardised evaluation protocols with realistic black-box threat models; certification formats readable by security managers.

Technical Priorities (High–Moderate)

T-R3—Cross-Dataset Generalisation Benchmarks: Mandatory out-of-distribution testing; contrastive and self-supervised pre-training targeting <5 pp accuracy drop.

T-R4—Distributed IDS Scalability: Benchmarking IDS pipelines on real cluster infrastructure (Kubernetes, Apache Flink); reporting throughput and latency under representative traffic loads.

Human-Centered Priorities

H-R1 (Critical)—RAG-XAI Comparative Evaluation: A small-scale but rigorous study ($n \geq 15$ SOC analysts) comparing SHAP-only vs. RAG-augmented explanations on triage speed, accuracy, and trust formation would provide the *minimum* empirical grounding currently absent for the RAG-XAI claim.

H-R2 (Critical)—Trust Calibration Longitudinal Study: Track analyst trust evolution from initial deployment through sustained use; identify trust collapse triggers.

H-R3 (High)—Usable Privacy Controls: Identify minimum transparency guarantees sufficient to overcome the privacy adoption barrier identified in the UTAUT2 study.

H-R4 (Moderate)—Alert Fatigue Mitigation: IDS interfaces with progressive disclosure and intelligent batching measured over $\geq$6-month deployments.

H-R5 (Critical)—Independent UTAUT2 Replication: The adoption findings in this survey rest on a single study [4] published in a venue with limited peer-review oversight. Independent replication using the published instrument on a probability-based sample (stratified by organisation size, industry sector, and geography) is essential to confirm or revise the adoption determinant hierarchy.

System-Level Priorities

S-R1 (Critical)—SOC Integration Pipelines: Open-source SIEM adapters for Splunk, Microsoft Sentinel, and IBM QRadar implementing Stage S8 continual learning with automated retraining triggers.

S-R2 (High)—Federated IDS Testbeds: Multi-organisation testbeds providing privacy-preserving production traffic evaluation.

S-R3 (High)—Sociotechnical Evaluation Infrastructure: Open-source toolkits implementing the Dual-Metric Framework; propose as IEEE/European Union Agency for Cybersecurity (ENISA)/NIST standards.

S-R4 (High)—Cluster Deployment Benchmarks: IDS pipeline benchmarks on representative cluster configurations (edge nodes + cloud aggregation) using the throughput–latency model (Equation (2)) as a baseline; standardised reporting of: (a) sustainable flows/s at the edge tier, (b) end-to-end alert-to-explanation latency, (c) federated communication overhead per aggregation round, and (d) Kubernetes pod scaling response time under traffic bursts. Benchmarks should target at least two traffic scales (10 Gbps campus and 40 Gbps data centre) and compare at least two stream-processing backends (Table 13).

6.2 Threats to Validity and Sensitivity Analysis

Internal Validity

Taxonomy coding subjectivity ($\kappa = 0.79$); quality score subjectivity in the 0–5 rubric; outcome conflation from paper-reported values without independent replication (selective reporting bias); the causal link between the meta-analysis and UTAUT2 results is observational, not experimentally established.

External Validity

English-only restriction excludes an estimated 15–20% of global cybersecurity output; publication bias under-represents negative results; recency bias favours high-citation preprints from 2022–2025.

Construct Validity

71% of papers use UNSW-NB15 or CIC-IDS2017, so meta-analysis statistics primarily characterise performance on these benchmarks. The human-factors axes (A1–A2) are derived from a single cross-sectional UTAUT2 study; their independence has not been tested through discriminant analysis, and they function as aspirational design requirements rather than classification axes (Section 4.2). The deployment rate $D = 0.32$ in the PAG framework derives from a general enterprise AI adoption report [6], not IDS-specific deployment data; the PAG sensitivity analysis (Table 8) demonstrates robustness to the weighting parameter but cannot compensate for imprecision in the underlying operationalisation. The UTAUT2 study [4] is published in a venue with limited peer-review oversight; the findings are treated as preliminary throughout this survey (see transparency note in Section 2.3).

Sensitivity Analysis

We assessed the robustness of the meta-analysis under three perturbations:

1. **Excluding arXiv preprints** ($n = 14$ removed): mean accuracy changes by <0.3 pp (97.1% $\rightarrow$ 97.4%); minority-class recall median unchanged at 0.648. Conclusions are insensitive to arXiv inclusion.
2. **Raising quality threshold** ($\geq 3/5 \rightarrow \geq 4/5$): corpus reduces to $n = 89$ papers; mean accuracy 97.4%; recall median 0.672. Recall collapse finding persists.
3. **Restricting to post-2019 papers** ($n = 98$): DL share increases to 52%; cross-dataset accuracy drop remains $\approx$17 pp. Temporal restriction does not qualitatively alter conclusions.

Reproducibility

The taxonomy coding rubric, quality scoring instrument, and paper-level extracted data are provided as supplementary material. The UTAUT2 survey instrument and anonymised dataset are available from the corresponding author upon reasonable request.

6.3 Theoretical and Practical Contributions

Theoretical contributions. The PAG conceptual framework (Definition 1) decomposes a phenomenon that practitioners observe into measurable generalisation and deployment components, with explicit sensitivity analysis. The Six-Axis Taxonomy provides a codeable design space spanning technical and human-deployment properties (with the caveat that A1–A2 are aspirational requirements, not classification axes). The Dual-Metric Framework operationalises PAG measurement.

Practical recommendations. *For IDS researchers:* routinely report minority-class recall and cross-dataset evaluation; publish negative results. *For system designers:* validate RAG-XAI empirically before making deployment claims; address privacy by design. *For organisations:* Privacy Concerns, not technical performance, are the primary deployment barrier; address data-handling transparency proactively. *For cluster computing practitioners:* use the analytical throughput–latency model

(Equation (2)) to size edge-cloud topologies; benchmark detection throughput alongside accuracy; consider Apache Flink for sub-10-ms latency IDS pipelines (Table 13). *For policymakers:* NIS2 explainability requirements align with the adoption evidence.

6.4 Limitations

This study has several limitations that should be considered when interpreting the findings. First, the systematic review is restricted to English-language publications, which excludes an estimated 15–20% of the global cybersecurity research output and may introduce a geographic and cultural bias toward Western institutional contexts. Second, the meta-analysis relies on paper-reported performance metrics without independent replication; selective reporting bias may inflate the aggregate accuracy figures and understate failure modes. Third, 71% of the reviewed studies evaluate on only two benchmark datasets (UNSW-NB15 and CIC-IDS2017), limiting the generalisability of the quantitative findings to broader network environments.

Fourth, the UTAUT2 adoption survey ($n = 300$) employs convenience-based sampling through professional cybersecurity channels, which may over-represent organisations with established AI adoption programmes and under-represent small enterprises or non-Western operational settings. Moreover, 84.3% of respondents assessed RAG-augmented IDS functionality based on scenario descriptions rather than direct hands-on experience, meaning the adoption constructs predominantly capture *hypothetical behavioural intentions* rather than actual usage behaviour. Fifth, the deployment rate proxy ($D = 0.32$) used in the PAG formalisation derives from a general enterprise AI adoption report rather than IDS-specific deployment data; while the sensitivity analysis demonstrates robustness to weighting parameter choices, the underlying operationalisation remains approximate.

Sixth, the causal link between technical performance gaps identified in the meta-analysis and the adoption barriers measured by the UTAUT2 study is observational, no experimental manipulation was performed to establish causality. Seventh, the RAG-augmented XAI architecture is proposed as a design direction supported by expert-panel ratings (6.4/7 projected usability) but lacks empirical validation from deployed production systems. Finally, the taxonomy coding, while achieving acceptable inter-rater reliability ($\kappa = 0.79$), involves inherent subjectivity in the quality scoring rubric, particularly for the human-factors axes (A1–A2), which function as aspirational design requirements rather than empirically validated classification dimensions. These limitations collectively suggest that the findings should be interpreted as indicative rather than definitive, and future work should prioritise longitudinal field studies with deployed IDS to validate the sociotechnical framework proposed herein.

7 Conclusion

This PRISMA-compliant systematic review of 153 empirical network-based IDS studies (2009–2025) and a UTAUT2 adoption survey ($n = 300$) investigates the Performance-Adoption Gap from a sociotechnical perspective. The meta-analysis

reveals three robust findings: benchmark accuracy does not predict operational performance (cross-dataset drop of 18 pp); minority-class recall collapse leaves critical attack types systematically undetected (median recall 0.648); and most DL models fail the sub-1-ms real-time deployment threshold.

The PAG conceptual framework (Equation (1)) decomposes this divergence into generalisation and deployment components, with a sensitivity analysis demonstrating that the qualitative conclusion is robust across weighting choices. The Six-Axis Taxonomy, the ten-stage architecture with pseudocode and analytical throughput–latency model, and the Dual-Metric Framework provide structured guidance for bridging technical and sociotechnical design.

RAG-augmented XAI is identified as a promising design direction with expert-panel support (6.4/7 projected usability) but no empirical validation from deployed systems. The most urgent contribution the community can make is a rigorous small-scale SOC study comparing SHAP-only vs. RAG-augmented explanations.

Closing the PAG requires parallel advances in technical engineering, sociotechnical design, and scalable distributed deployment, a research agenda this survey charts with explicit caveats on its evidence base.

References

[1] IBM Security. Cost of a Data Breach Report 2024. Armonk, NY: IBM Corporation; 2024.

[2] Morgan S. Cybersecurity jobs report: 3.5 million unfilled positions in 2025. Cybercrime Magazine. 2023 Nov.

[3] Ponemon Institute. The cost of cybersecurity alert overload. Ponemon Institute Research Report. 2022.

[4] Ren Y. Data science and machine learning for cyber intrusion detection: a systematic review. Mach Learn Res. 2026;11(1):8–21. https://doi.org/10.11648/j.mlr.20261101.12

[5] Apruzzese G, et al. The role of machine learning in cybersecurity. Digit Threats Res Pract. 2023;4(1).

[6] O'Reilly Media. AI adoption in the enterprise 2024. O'Reilly Industry Research Report. 2024.

[7] Sasse MA, Fléchais I. Usable security: why do we need it and how do we get it? In: Security and Usability. Sebastopol: O'Reilly; 2005. p. 13–30.

[8] Amershi S, et al. Software engineering for machine learning: a case study. In: Proc IEEE/ACM Int Conf Softw Eng Softw Eng Pract (ICSE-SEIP); 2019. p. 291–300.

[9] Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutor. 2016;18(2):1153–1176.

[10] Liao HJ, et al. Intrusion detection system: a comprehensive review. J Netw Comput Appl. 2013;36(1):16–24.

[11] Ferrag MA, et al. Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study. J Inf Secur Appl. 2020;50:102419.

[12] Abdulganiyu OH, et al. A systematic literature review on network intrusion detection based on deep learning. Expert Syst Appl. 2023.

[13] Arrieta AB, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82–115.

[14] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Adv Neural Inf Process Syst; 2017. p. 4765–4774.

[15] Marino DL, Wickramasinghe CS, Manic M. An adversarial approach for explainable AI in intrusion detection systems. In: Proc IECON; 2018. p. 3237–3243.

[16] Arreche O, et al. XAI meets IDS: integrating SHAP-based explanations into XGBoost intrusion detection. IEEE Trans Inf Forensics Secur. 2025.

[17] Lewis P, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Adv Neural Inf Process Syst; 2020. p. 9459–9474.

[18] Houssel P, et al. Towards explainable network intrusion detection using large language models. arXiv:2408.04342. 2024.

[19] Blefari F, et al. CyberRAG: an agentic RAG cyber attack classification tool. arXiv:2507.02424. 2025.

[20] Zurko ME, Simon RT. User-centered security. In: Proc New Secur Paradigms Workshop; 1996. p. 27–33.

[21] Bada M, Sasse AM, Nurse JRC. Cyber security awareness campaigns: why do they fail to change behaviour? arXiv:1901.02672. 2019.

[22] Venkatesh V, et al. User acceptance of information technology: toward a unified view. MIS Q. 2003;27(3):425–478.

[23] Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the unified theory. MIS Q. 2012;36(1):157–178.

[24] Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors. 2004.

[25] Trist E. The evolution of sociotechnical systems as a conceptual framework and as an action research program. In: Perspectives on Organization Design and Behavior. New York: Wiley; 1981. p. 19–75.

[26] Baxter G, Sommerville I. Socio-technical systems: from design methods to systems engineering. Interact Comput. 2011;23(1):4–17.

[27] Page MJ, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

[28] Moustafa N, Slay J. UNSW-NB15: a comprehensive dataset for network intrusion detection systems. In: Proc MilCIS; 2015.

[29] Sharafaldin I, et al. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proc ICISSP; 2018.

[30] Koroniotis N, et al. Towards the development of realistic botnet dataset in the IoT for network forensic analytics: Bot-IoT dataset. Future Gener Comput Syst. 2019.

[31] Moustafa N. A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets. Sustain Cities Soc. 2021.

[32] Neto ECP, et al. CICIoT2023: a real-time dataset and benchmark for large-scale attacks in IoT. Sensors. 2023.

[33] Ferrag MA, et al. Edge-IIoTset: a new comprehensive realistic cyber security dataset of IoT and IIoT applications. IEEE Access. 2022.

[34] Dadkhah S, et al. CICIoMT2024: a benchmark dataset for assessing IoMT device security. Internet Things. 2024.

[35] Mishra P, et al. A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Commun Surv Tutor. 2019.

[36] Gao X, et al. An adaptive ensemble machine learning model for intrusion detection. IEEE Access. 2019.

[37] Dhaliwal SS, et al. Effective intrusion detection system using XGBoost. Information. 2018.

[38] Kilincer IF, et al. Machine learning methods for cyber security intrusion detection: datasets and comparative study. Comput Netw. 2021.

[39] Choudhary S, Kesswani N. Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 datasets using deep learning with LSTM. Procedia Comput Sci. 2020.

[40] Rahman MM, et al. A survey on intrusion detection system in IoT networks. Cyber Secur Appl. 2025.

[41] Saba T, et al. Anomaly-based intrusion detection system for IoT networks through deep learning model. Comput Electr Eng. 2022.

[42] Alkhatib K, Mohammed S. Transformer-based network intrusion detection for IoT using edge computing. IEEE Internet Things J. 2024.

[43] Nwosu EC, et al. Graph-based network intrusion detection for heterogeneous data. IEEE Trans Netw Serv Manag. 2022.

[44] Zhou J, et al. Graph neural networks: a review of methods and applications. AI Open. 2020.

[45] Mothukuri V, et al. A survey on security and privacy of federated learning. Future Gener Comput Syst. 2021.

[46] Liu X, et al. Survey of federated learning in intrusion detection. J Parallel Distrib Comput. 2025.

[47] Ferrag MA, et al. DeepCoin: a novel deep learning and blockchain-based energy exchange framework for smart grids. IEEE Trans Eng Manag. 2021;68(1):1999–2012.

[48] Ferrag MA, et al. SecureBERT and LLAMA 2 empowered control area network intrusion detection and classification. EAI Endorsed Trans Creat Technol. 2024.

[49] Rajagopal S, et al. A stacking ensemble for network intrusion detection using heterogeneous datasets. Secur Commun Netw. 2020.

[50] Tama BA, et al. TSE-IDS: a two-stage classifier ensemble for anomaly-based intrusion detection system. IEEE Access. 2019.

[51] Sundaramurthy SC, et al. Turning contradictions into innovations or: how we learned to stop whining and embrace the complexity of the security operations center. In: Proc SOUPS; 2014.

[52] Tavallaee M, et al. A detailed analysis of the KDD CUP 99 data set. In: Proc IEEE Symp Comput Intell Secur Def Appl (CISDA); 2009. p. 1–6.

[53] Alhajjar E, Maxwell P, Bastian N. Adversarial machine learning in network intrusion detection systems. Expert Syst Appl. 2021;186:115782.

[54] Sarhan M, et al. Towards a standard feature set for network intrusion detection system datasets. Mob Netw Appl. 2022;27:357–370.

[55] Ring M, et al. A survey of network-based intrusion detection data sets. Comput Secur. 2019;86:147–167.

[56] Karimov J, et al. Benchmarking distributed stream data processing systems. In: Proc IEEE Int Conf Data Eng (ICDE); 2018. p. 1507–1518.

[57] Preuveneers D, et al. Resource usage and performance trade-offs for machine learning models in smart environments. Sensors. 2020;20(4):1176.

Appendix A Empirical UTAUT2 Study: Extended RAG Cybersecurity Adoption

A.1 Study Overview

This study investigates adoption of RAG-enhanced AI cybersecurity assistance systems among security professionals through an extended UTAUT2 framework [23] incorporating three RAG-specific constructs—Trust & Accuracy (TR), Data Privacy Concerns (PR1), and Low Risk Belief (PR2)—validated through a survey of $n = 300$ cybersecurity professionals (68.7% technical, 31.3% non-technical) and 39 qualitative interviews (January–August 2024).

Sample demographics. Role distribution: 38% SOC operations, 24% threat intelligence, 21% incident response, 17% other security roles. Mean experience: 8.4 years (SD = 4.2). Organisational size: 42% large enterprise (>5,000 employees), 31% mid-size (500–5,000), 27% small (<500). Geography: 61% North America, 23% Europe, 16% Asia-Pacific. Instrument: 63 items, 7-point Likert scale; expert panel validated ($n = 12$), pilot tested ($n = 45$); Cronbach's α: 0.881–0.924.

A.2 Key Results

The extended model explains 56.4% of variance in Behavioural Intention, a 15.2 pp increment over the baseline UTAUT2 model ($\Delta F(3, 289) = 33.72$, $p < 0.001$).

A.3 Sample, Methodology, and Limitations

Sampling strategy. Participants were recruited through a combination of purposive and snowball sampling via three channels: (1) professional cybersecurity organisations (ISACA and $(ISC)^2$ chapter mailing lists); (2) LinkedIn professional groups focused on SOC operations and threat intelligence; and (3) direct outreach to cybersecurity teams at six organisations (two Fortune 500 companies, two government agencies, two managed security service providers). This convenience-based sampling strategy limits generalisability to the broader population of security professionals; results may over-represent organisations with established AI adoption programmes and under-represent small enterprises or non-Western contexts.

RAG experience. Of the 300 respondents, 47 (15.7%) reported direct hands-on experience with RAG-enhanced security tools (e.g., ChatGPT-based SOC assistants, vendor-provided AI triage tools with retrieval components). The remaining 253

Table A1 UTAUT2 regression results ($n = 300$ security professionals). BI = Behavioural Intention; PE = Performance Expectancy; EE = Effort Expectancy; FC = Facilitating Conditions; SI = Social Influence; n.s. = not significant.

Construct	β	**SE**	p	**Interpretation**
Privacy Concerns (PR1)	−0.246	0.047	<0.001	Strongest negative predictor; primary adoption barrier
Performance Expectancy (PE)	+0.235	0.058	<0.001	Strongest positive predictor; productivity drives adoption
Trust & Accuracy (TR)	+0.223	0.051	<0.001	Direct predictor AND partial mediator of PE→BI (indirect effect = 0.0862, 95% CI: [0.0514, 0.1247])
Low Risk Belief (PR2)	+0.210	0.049	<0.001	Perceived organisational safety facilitates adoption
Effort Expectancy (EE)	+0.156	0.051	<0.01	Stronger for technical users ($\beta = 0.187$) vs. non-technical ($\beta = 0.098$)
Facilitating Conditions (FC)	+0.134	0.049	<0.05	Infrastructure and training availability
Social Influence (SI)	+0.089	0.047	n.s.	Non-significant overall; stronger for novices ($\beta = 0.201$)
Model Fit: $R^2 = 0.564$, Adj $R^2 = 0.548$		$F(10, 289) = 35.82^{***}$; $\Delta R^2 = 0.152$ vs. UTAUT2-only ($R^2 = 0.412$)		

$^{***}p < 0.001$. Bootstrap mediation: 5,000 resamples. $\Delta F(3, 289) = 33.72$, $p < 0.001$ for RAG constructs.

(84.3%) responded based on detailed scenario descriptions of RAG-augmented IDS functionality provided in the survey instrument (including screenshots of a prototype RAG-XAI alert interface). Therefore, the study predominantly measures *hypothetical adoption intentions* rather than actual adoption behaviour for RAG-specific constructs. This distinction is critical: behavioural intention is a necessary but insufficient predictor of actual adoption, and the gap between stated intentions and realised behaviour is well-documented in technology adoption research.

CFA fit indices. Confirmatory Factor Analysis model fit: CFI= 0.923 (below the commonly recommended ≥0.95 threshold for good fit; we note that CFI values between 0.90 and 0.95 are considered acceptable in applied research with complex models [23]); RMSEA= 0.057 (90% CI: [0.048, 0.066], within the acceptable range); SRMR= 0.064; AVE: 0.623–0.749; HTMT ratios: 0.312–0.687; VIF average= 1.12.

Qualitative component. The 39 semi-structured interviews (stratified by role: 15 Tier 1 analysts, 12 Tier 2 investigators, 7 Tier 3 threat hunters, 5 managers) were analysed using thematic analysis (Braun & Clarke, 2006). Three themes emerged that contextualize the quantitative findings: (1) *Explanation verification need*—analysts reported that unexplained alerts are routinely deprioritised regardless of severity score, supporting the Trust & Accuracy pathway ($\beta = +0.223$); (2) *Privacy anxiety*— multiple respondents described institutional resistance to cloud-based AI tools due to data sovereignty concerns, consistent with the Privacy Concerns barrier ($\beta = -0.246$); (3) *Workflow integration friction*—analysts described abandoning tools that required context-switching between the SIEM console and separate AI interfaces, supporting the Effort Expectancy finding ($\beta = +0.156$).

Relationship to this survey. The UTAUT2 study was authored by the first author of the present survey and published separately [4]. The present survey relies on the published findings but cannot independently verify them. The survey instrument,

anonymised dataset, and full CFA output tables are available from the corresponding author to enable independent replication. We treat the UTAUT2 findings as preliminary evidence pending such replication (see also the transparency note in Section 2.3).

A.4 Implications for IDS Design

(1) Address Privacy Concerns proactively through on-premises deployment and transparent data-handling. (2) Emphasise citation transparency—Trust & Accuracy mediates adoption. (3) Communicate risk mitigation actively—management needs compliance certificates, not model accuracy reports. (4) Maintain performance focus—privacy controls enable adoption, but demonstrated productivity sustains it. (5) Tailor by role—technical users need efficient integration; novices respond to social endorsement. Full instrument and dataset available from the corresponding author.

Appendix B IDS Techniques Comparison

Table B2 IDS techniques comparison: accuracy, minority-class recall, explainability, and cluster deployment readiness. VAE = Variational Autoencoder; GRU = Gated Recurrent Unit; FedAvg = Federated Averaging; DP = Differential Privacy.

Model	Dataset	Accuracy (%)	Min-Class Recall	Explainability	Deployment / Cluster Readiness
Naïve Bayes	NSL-KDD	92–95	Low (60–70%)	Intrinsic (feature priors)	Lightweight; edge-suitable; low cluster resource
Logistic Regression	NSL-KDD / UNSW-NB15	90–93	Low (55–65%)	Intrinsic (coefficients)	Lightweight; real-time; horizontal scale trivial
Decision Tree	KDD'99 / UNSW-NB15	95–97	Moderate (70–80%)	Intrinsic (rule paths)	Very lightweight; IoT edge nodes
Random Forest	UNSW-NB15 / CIC-IDS2017	97–99	Moderate (72–82%)	Post-hoc (feature importance)	Moderate; batch or streaming; good cluster throughput
SVM	NSL-KDD / UNSW-NB15	96–98	Moderate (68–78%)	Post-hoc (SHAP)	Moderate; latency concern at line rate
XGBoost	CIC-IDS2017 / CSE-CIC-IDS2018	98–99	Moderate (75–85%)	Post-hoc (SHAP/LIME)	High; cloud-ready; SIEM-integrable; Spark-compatible
CNN	CIC-IDS2017 / UNSW-NB15	97–99	Moderate (73–83%)	Post-hoc (Grad-CAM)	Cloud; GPU cluster required
LSTM / Gated Recurrent Unit (GRU)	KDD'99 / UNSW-NB15	97–99	Moderate (70–80%)	Post-hoc (attention)	Cloud; high memory; sequential; not line-rate
CNN–LSTM Hybrid	CIC-IDS2017 / Bot-IoT	98–99.5	Moderate-High (78–88%)	Hybrid (attention+SHAP)	Cloud/edge-cloud split; best accuracy–efficiency
Autoencoder / VAE	UNSW-NB15 / CIC-IDS2017	95–98 (anomaly)	Low-Mod (60–75%)	Reconstruction error	Unsupervised; zero-day capable; moderate cluster cost
Transformer / Attention	CIC-IDS2017 / Bot-IoT	98–99.5	High (80–90%)	Attention maps	Cloud cluster; quadratic complexity; offload from edge
GNN-based IDS	UNSW-NB15 / custom graphs	97–99	High (80–88%)	Graph attention	Cloud cluster; lateral movement; graph database integration
Federated (FedAvg+DP)	TON_IoT / Edge-IIoTset	94–97	Moderate (68–78%)	Federated SHAP	Multi-org cluster; privacy-preserving; horizontal federations
LLM/RAG-based	Logs + CVE/NVD	N/A (reasoning)	Context-dependent	RAG citations (highest)	SOC triage layer only; high latency; not line-rate