AI-Based Hybrid Anomaly Detection and Behavioral Threat Response Systems: A Comprehensive Review of Advances, Challenges, and Future Directions

Pankaj Deshmukh[1], Saiz Momin[2], Kshitij Thakkar[3], Tarun Kandarpa[4]

[1,2,3,4]Artificial Intelligence & Data Science
K J Somaiya Institute of Technology, Sion
Mumbai, India
[1]pankaj@somaiya.edu, [2]saiz.momin@somaiya.edu
[3]kshitij.t@somaiya.edu, [4]tarun.k@somaiya.edu

Abstract

Intrusion Detection Systems (IDS) have progressed from signature-based models to hybrid frameworks powered by artificial intelligence, incorporating machine learning and extensive language models. This review brings together research on AI-based systems for finding unusual behavior and responding to threats, with a focus on hybrid architectures, explainable AI (XAI), and reasoning based on LLMs. The paper talks about existing problems including false positives, lack of explainability, and scalability. It also talks about trends in adaptive risk prioritization, federated IDS deployment, and semantic enrichment using MITRE ATT&CK frameworks.

Keywords: Intrusion Detection System (IDS), Anomaly Detection, Explainable AI, Large Language Models, Cybersecurity, Hybrid IDS, MITRE ATT&CK, Threat Response

1. Introduction

Modern networks are subjected to an onslaught of cyberattacks that consistently exploit novel vulnerabilities and strategies [8], [9]. Intrusion Detection Systems (IDS) continue to be a critical defense layer, monitoring network or host activity for malicious behavior [5]. Traditional NIDS, such as Snort or Suricata, employ signature-based detection, which involves comparing incoming traffic to a database of recognized attack patterns[1]. Signature engines are highly effective in detecting known threats; however, they are unable to identify zero-day or novel attacks that lack corresponding signatures [5], [12]. Consequently, covert or polymorphic threats frequently evade detection. This constraint is especially severe due to the prevalence of encrypted or obfuscated payloads by adversaries, which only permit the observation of traffic metadata [8].

Anomaly-based IDS develop a statistical model of "normal" behavior during a training phase and identify deviations as probable intrusions. Machine learning (ML) and deep learning (DL) techniques, including autoencoders, LSTM networks, and convolutional architectures, have been utilized to represent standard network flows or host events [5]. These methodologies can identify previously unrecognized attacks without specific signatures. In dynamic cloud environments, deep learning models (CNNs, LSTMs) have achieved approximately 98–99% accuracy in identifying distributed attacks [12]. Nonetheless, anomaly detectors typically experience elevated false-positive rates, as differentiating benign deviations from genuine attacks is difficult. Excessive notifications can inundate security analysts, leading to alert fatigue and diminishing confidence in the system.

Hybrid IDS architectures have emerged to reconcile these trade-offs. Such systems integrate signature and anomaly detection within a unified architecture [5]. The signature protocols swiftly obstruct recognized threats, but a machine learning-based behavioral model serves as a safeguard against unidentified anomalies. Recent attempts, for instance, combine Suricata's rules with a TensorFlow/Keras autoencoder trained on standard network flows[13]. Alerts from both systems are interconnected to enhance detection efficacy: recognized assaults are identified through signatures, while new or covert ones are detected via behavioral anomalies. This multilayer strategy seeks to minimize missed detections while avoiding an increase in false positives.

Simultaneously, developments in Explainable AI (XAI) and Large Language Models (LLMs) are enhancing the usability of Intrusion Detection Systems (IDS). The opacity of numerous ML/DL classifiers constrains analyst confidence and situational awareness. XAI methodologies, such as SHAP and LIME, have been employed to deliver feature-level elucidations for IDS determinations [1], [19]. Recently, large language models (LLMs) like as GPT-4 and Llama have been suggested as AI agents for intrusion detection systems (IDS); they can process alerts, network and host context (e.g., MITRE ATT&CK data), and produce comprehensible incident reports or explanations. The "IDS-Agent" architecture employs a large language model in a reasoning loop to identify IoT network data and elucidate its conclusions, attaining an approximate F1-score of 0.97 on benchmarks. The “eX-NIDS” architecture enhances identified flows with threat intelligence context and instructs an LLM to generate comprehensive explanations, resulting in a performance improvement of around 20% over a baseline . These capabilities are poised to significantly expedite analysis by converting raw alarms into comprehensible narratives [2].

This research examines the hybrid AI-based Intrusion Detection System landscape, integrating signature rules, anomaly detection through machine learning and deep learning, explainable AI, and context enrichment powered by large language models. We examine cutting-edge approaches and performance, compare exemplary systems, and underscore existing problems (false positives, interpretability, scalability, human-in-the-loop). We delineate prospective avenues including adaptive and continuous learning, federated intrusion detection systems for decentralized contexts, semantic enhancement (e.g., augmenting alarms with ATT&CK context), and artificial intelligence-driven risk ranking.

2. Literature Survey

A. Signature Based IDS

Signature-based Intrusion Detection Systems have been fundamental in network security. Tools such as Snort and Suricata utilize pattern-matching engines to examine packets or flows for byte sequences, protocol irregularities, or other indicators specified in a signature database. These systems are proficient at identifying recognized exploits (e.g., certain buffer overflow signatures), exhibiting minimal false positives for such patterns. Nevertheless, by design, they are incapable of detecting attacks that do not conform to established signatures. When attackers employ new exploits or subtly altered variations, a signature engine will only permit those occurrences passively. Furthermore, the upkeep of current signature libraries is an ongoing endeavor: new protocols must be developed and disseminated for evolving threats [5], resulting in intervals of susceptibility. As indicated in [5], this reactive characteristic implies that signature-based IDS "cannot detect attacks that do not correspond to any existing signature" . Consequently, although signatures are essential, their constraints highlight the necessity for supplementary methods.

B. Anomaly Based Detection

Anomaly detection approaches were developed to address the limitations of signatures. Rather than relying on set patterns, these strategies establish a baseline of typical network or system activity and identify statistical anomalies. Initial methodologies employed statistical thresholds, however contemporary techniques utilize machine learning and deep learning. An autoencoder network can be trained on standard traffic flows to reconstruct inputs; a significantly elevated reconstruction error on a flow indicates an anomaly[13]. LSTM and other recurrent networks may effectively simulate time-series patterns in traffic or system-call sequences, thereby capturing temporal correlations in normal activity [6], [17]. Convolutional Neural Networks (CNNs) have been utilized on time-windowed feature matrices of network metrics, with great accuracy in specific domains[12].

These AI-based detectors exhibit potential, with studies indicating detection accuracy of approximately 98–99% on conventional datasets or cloud traffic for deep-learning models [12], [13]. Sowmya and Anita (2023) examine 72 AI-IDS research and determine that machine, deep, and ensemble learning significantly enhance detection accuracy compared to conventional approaches. In a particular practical evaluation [7], [18], a compact CNN utilized in a federated IoT IDS attained approximately 98% accuracy with minimal inference delay, although a more intricate CNN+BiLSTM achieved around 99% accuracy at a higher expense[12]. This performance indicates that ML/DL may detect intricate, multi-faceted attack patterns, such as DDoS flows, that signatures fail to recognize.

Nonetheless, anomaly detectors encounter considerable difficulties. The primary concern is elevated false-alarm rates, as innocuous fluctuations (e.g., flash crowds, traffic surges) frequently activate warnings. Adjusting sensitivity poses challenges: establishing low thresholds captures numerous attacks but inundates analysts with false positives; elevating thresholds overlooks nuanced incursions. The literature consistently underscores this trade-off. One survey indicates that "anomaly-based IDS often encounter elevated false-positive rates, which can inundate SOCs with non-actionable alerts" [5], [18]. In practice, minimizing false positives while preserving memory continues to be a prominent research focus. Additional concerns encompass the necessity for representative training data, as unsupervised models presuppose a consistent "normal" setting, and the challenges associated with categorizing and assessing realistic traffic.

C. Hybrid IDS Architectures

Hybrid Intrusion Detection Systems integrate signature and anomaly detection engines to utilize the advantages of both methodologies. In a layered hybrid, incoming traffic is concurrently evaluated against static rules (efficient and accurate for recognized risks) and analyzed using a behavioral model (detecting unfamiliar patterns) [13], [15]. This method seeks to reduce blind spots: recognized malware activates signatures, whereas zero-day or stealthy payloads that bypass signatures can be detected through deviations. Recent prototypes illustrate this strategy: one version integrates Suricata's signature engine with a TensorFlow autoencoder trained on benign NetFlow features[21]. When Suricata triggers an alert or the autoencoder surpasses an anomaly threshold, the system produces an associated alert. The hybrid engine effectively "balances precision against known threats with the capacity to recognize atypical behavioral patterns that signatures alone cannot detect." [22]

Moreover, hybrid designs often integrate both network-level and host-level monitoring. For instance, host-based agents can monitor system calls or process activities, feeding an LSTM model, while the network engine analyzes packet flows. Sworna et al. highlight that considering system calls as "language tokens" and applying NLP models (e.g. LSTM+CNN) produces extraordinarily high detection rates (up to 99.9%) for host anomalies [6]. By unifying multiple data sources, hybrids achieve more holistic visibility. In brief, hybrid systems can identify multi-vector attacks and insider threats that can get past single-layer solutions.

D. Explainable AI (XAI) in IDS

As machine learning and deep learning models become increasingly ubiquitous, interpretability has emerged as essential for analyst confidence and regulatory adherence. XAI methodologies seek to elucidate obscure classifiers. In Intrusion Detection Systems, this entails disclosing the specific attributes or flows that contributed to a threat determination. Mane and Rao (2021) developed a XAI framework for IDS by employing a deep neural network on the NSL-KDD dataset, subsequently utilizing SHAP, LIME, contrastive explanations, and rule extraction to elucidate its conclusions [1]. Their approach attained approximately 82% identification accuracy while offering human-interpretable justifications using SHAP values that denote feature significance. This transparency enables analysts to verify and refine models.

Recent systems advance Explainable Artificial Intelligence (XAI) further. An "XAI-IDS" system combines DNNs with SHAP/LIME to provide local and global explanations, achieving 82% accuracy with complete interpretability[1]. The eX-NIDS architecture advances by employing LLMs to analyze alerts: it initially enhances harmful flow records with contextual threat intelligence, subsequently prompting a model such as GPT-4 to generate an explanation for the malicious classification of the flow [2]. This augmented prompt technique produces more precise explanations than a standard prompt baseline, thereby transforming the LLM into a sophisticated explainer.

Explainability pertains not only to model outputs but also to incident narratives. Large Language Models can convert raw warnings and logs into comprehensible summaries. For example, the IDS-Agent's workflow incorporates a step where a large language model generates sequential reasoning for a classification, allowing it to elucidate the rationale behind categorizing traffic as benign or malicious. These XAI/LLM hybrids enhance situational awareness. The literature emphasizes that interpretable AI is essential for implementation in security operations, enabling analysts to trust and act upon AI recommendations.

E. LLM Integration in IDS

Beyond explainability, Large Language Models are being embedded directly into IDS as analytic engines. Recent research demonstrates LLMs' utility in contextualization and automated response. For vulnerability management, Rafiey and Namadchian (2025) used GPT-3 to map CVE descriptions to MITRE ATT&CK techniques, achieving comparable results to fine-tuned BERT models at much lower cost [10]. In intrusion detection, Li et al. (2024) introduced IDS-Agent, an LLM-powered agent that actively orchestrates data preprocessing, classification, and explanation. It maintains memory and uses iterative prompting to refine its verdict [4]. IDS-Agent outperformed traditional ML and basic LLM-baseline models on IoT benchmarks (F1-scores ~0.97 and 0.75) and even detected zero-day attacks with ~61% recall [4].

Similarly, Houssel et al. (2025) showed that an LLM with enriched prompts can yield high-quality explanations for NIDS alerts [2]. They focus on generative explanation quality rather than raw classification accuracy, but this context-markup approach significantly enhances clarity. In practice, such systems can automate the “triage” step of security operations: transforming raw IDS logs into structured, prioritized incident reports. The project report motivating this review uses an LLM with a FAISS-indexed ATT&CK knowledge base to enrich alerts, generating concise, actionable summaries for SOC teams [20].

In summary, while LLMs alone may not yet match specialized IDS accuracy, they excel at correlation and narrative. By linking alerts to threat intelligence, summarizing chain-of-events, and suggesting mitigations, LLMs reduce analyst workload. Early results, such as IAM, show over 60% F1 in linking vulnerabilities to attack techniques. As LLMs improve, their role in IDS will likely expand to real-time incident analysis and even conversational interfaces for security analysts.

3. Comparative Table of Key IDS Approaches

“XAI-IDS” employs deep neural networks alongside explanation tools (SHAP/LIME), attaining approximately 82% accuracy while ensuring complete interpretability. eX-NIDS [1] integrates flows identified by an NIDS with an LLM to produce natural-language explanations, resulting in an enhancement of explanation quality by more than 20% due to the improved prompt. IDS-Agent is a framework based on large language models for the Internet of Things, demonstrating approximately 97% F1 score in distinguishing between benign and malicious traffic [4], [16]. The surveys indicate approximately 99–99.9% detection rates for ensemble and deep models on public datasets. An "AI-IDS for Cloud" attained 98.5% accuracy utilizing CNNs on cloud DDoS data. Ultimately, Olatunde's AI-driven engine, utilizing reinforcement learning, markedly reduced the Mean Time to Detect and Respond inside an enterprise risk framework. These works demonstrate the trade-offs of model complexity, interpretability, and practical implementation.

Framework / Study	Model / Technique	Key Performance	Use Case
XAI-IDS (Mane et al.)	Deep Neural Network + XAI (SHAP, LIME)	~82% accuracy with full interpretability	Transparent NIDS with human-in-the-loop validation
eX-NIDS (Houssel et al., 2025)	LLMs (Llama 3, GPT-4) with prompt augmentation	>20% improvement over baseline explainer	Enhanced explanations for malicious network flows

IDS-Agent (Li et al., 2024)	LLM Agent (GPT-4o) with memory and reasoning	0.97 F1 (ACI-IoT); 0.75 F1 (CIC-IoT); 0.61 recall on zero-days	Explainable IoT IDS with multi-step reasoning
Survey: AI-based IDS (Sowmya et al., 2023)	ML/DL ensemble methods	Up to 99.1% accuracy reported	Comprehensive review of AI-based IDS techniques
Survey: NLP for HIDS (Sworna et al., 2023)	LSTM, CNN on system call sequences	Up to 99.9% detection accuracy	Host-based anomaly detection via sequence modeling
AI-IDS for Cloud (Smith & Kevin, 2025)	CNN, LSTM models	~98.5% accuracy (cloud DDoS detection)	High-volume traffic analysis in cloud environments
Risk Prioritization (Olatunde, 2024)	Predictive analytics + Reinforcement Learning	83% reduction in MTTD; 87.5% reduction in MTTR	Asset-aware threat scoring in enterprise SOC

Table 1. Comparative analysis of different frameworks

4. Challenges

False Positives and Alarm Fatigue: Anomaly detectors frequently produce numerous false alerts due to the significant variability of typical traffic. Elevated false-positive rates undermine trust and impose a strain on analysts. Achieving equilibrium between sensitivity and specificity necessitates dynamic thresholds and contextual filtration. Aliyu et al. revealed that an LSTM-based continuous learner attained 99.94% accuracy and a remarkably low false-positive rate of 0.06% on NSL-KDD, but real-world networks exhibit greater complexity. False positives continue to be a significant challenge necessitating improved feature selection, model calibration, or feedback-informed learning.

Explainability and Trust: Deep models and large language models are potent yet frequently inscrutable. Lysenko et al. assert that AI must not only identify risks but also elucidate decisions to humans. [9]. In the absence of interpretability, analysts may lack confidence in an Intrusion Detection System or be incapable of auditing its warnings. XAI approaches like as SHAP and LIME, along with attention visualization, alleviate this issue; nevertheless, they introduce complexity and may not be scalable to high traffic levels. Furthermore, LLM explanations may appear credible yet be factually inaccurate (“hallucinations”), so engendering new trust concerns. Ensuring that explanations adequately represent the model's logic remains a persistent research challenge.

Scalability and Deployment: Modern IT environments include high-speed networks, cloud services, and IoT devices. Intrusion Detection Systems must be scalable across these varied areas. Distributed or federated methodologies are advantageous: by training localized models on each segment (particularly in privacy-sensitive IoT environments) and subsequently combining them, scalability is enhanced. Federated IDS presents problems related to heterogeneity, as devices exhibit significant variability in data and computational capabilities, as well as communication overhead. The MDPI work by Albanbay et al. shown that a compact CNN can operate on Raspberry Pi edge devices with approximately 98% accuracy; nevertheless, actual implementations must manage several nodes and imbalanced data. Overseeing model updates and synchronization while safeguarding sensitive data is complex.

Human AI Collaboration: Despite automation, human analysts continue to be indispensable. Intelligent Decision Systems must be engineered for optimal human–AI collaboration. This encompasses explicit visual representations, prioritization of notifications, and systems for analyst input. Numerous studies indicate that AI-augmented pipelines should complement rather than supplant human involvement. Integrating AI technologies into SOC operations presents challenges: excessive automation can undermine analysts' authority, whilst little automation does not reduce burden. The socio-technical concern of depending on LLM-generated reports necessitates the establishment of methods to validate and enhance AI outputs. Training and trust-building are essential; for instance, XAI-driven IDS seek to establish confidence by elucidating the rationale behind a model's alert selection.

5. Future Directions

Future prospects indicate numerous possible avenues to rectify existing deficiencies:
Adaptive continuous learning: Instead of static models, IDS should perpetually learn from new data and analyst input. Aliyu et al. offer an LSTM-based "self-adaptive" Intrusion Detection System (IDS) that retrains on incoming logs via a blockchain for distributed trust. Likewise, reinforcement-learning-based intrusion detection systems (e.g., DRL-IDS for SDN) modify policies in real time. Feedback loops in which analysts identify false positives or overlooked risks might enhance anomaly thresholds. Meta-learning and online learning techniques may facilitate the rapid adaptation of Intrusion Detection Systems to evolving traffic patterns (concept drift).

Federated and Edge IDS:As networks grow increasingly distributed (cloud-edge, IoT), federated learning (FL) emerges as a critical domain. Federated Learning permits each device or site to train a local model and disseminate just model updates, so safeguarding data privacy. Recent experiments validate the feasibility: in a simulated IoT network comprising up to 150 devices, a federated CNN attained around 98% accuracy on CICIoT2023. Future research will investigate the optimization of federated protocols for anomaly detection, the management of diverse data distributions, and the development of lightweight model architectures for resource-limited sensors. Semantic federated learning, which involves exchanging feature representations of network events instead of raw data, could enhance both privacy and efficiency.

Semantic Enrichment (Knowledge Graphs & ATT&CK): Enhancing alerts with structured threat intelligence, beyond mere abnormalities, is advantageous. Mapping IDS output to MITRE ATT&CK methods helps facilitate triage. The LLM-based correlation of CVEs to ATT&CK demonstrates initial success. Incorporating knowledge graphs, MITRE-based ontologies, and more semantic data into IDS helps enhance context-awareness. A hybrid system may autonomously annotate an alert with probable adversary strategies, impacted assets, or suggested mitigations, employing AI to integrate log data with external databases.

Risk-Based Prioritization: Not all signals possess equal significance. Future IDS must integrate risk scoring to contextualize threats based on business impact. Olatunde (2024) shown that AI-powered risk engines may associate threat severity with asset criticality, significantly reducing response times[32].

Building upon this, IDS might ascertain which anomalies are most probable to evolve into incidents, prioritizing them accordingly. Integrating with governance frameworks, such as compliance metrics, could customize warnings to align with an organization's priorities.

Innovations in Human–AI Interfaces: Innovations in human-centered AI, including conversational interfaces that enable analysts to question the IDS using natural language, are forthcoming. An analyst may query an IDS-agent: "Provide all high-severity alerts concerning file uploads from the past hour," and obtain both data and elucidations. LLM chat interfaces or augmented reality dashboards could enhance the actionability of IDS outputs. Investigations into explainable reinforcement learning and trust calibration will enhance collaboration.

Regulatory and Ethical Considerations: As AI assumes an increasingly significant role in security, monitoring is important. Preliminary research indicates that most existing cyber regulations fail to address AI-specific occurrences, such as algorithmic bias and model failure. Future IDS research must address privacy (e.g., differential privacy-preserving machine learning), fairness, and transparency compliance, particularly when LLM records encompass sensitive material.

6. Conclusion

The study examines advancements in AI-driven hybrid Intrusion Detection Systems (IDS) that integrate signature-based, anomaly-based, and explainable AI (XAI) methodologies, augmented by large language models (LLMs). These systems offer enhanced, stratified protections and augment analysts' comprehension of dangers. Despite achieving over 95% detection accuracy, existing models encounter obstacles like elevated false positive rates, restricted explainability, and scaling issues in IoT and cloud environments. The document underscores the necessity for adaptive, self-learning, and federated Intrusion Detection System architectures that incorporate semantic enrichment and risk prioritization. The future of IDS ultimately resides in intelligent, interpretable systems capable of contextual learning, decision explanation, and successful collaboration with human analysts.

References

1. S. Mane and D. Rao, "Explaining network intrusion detection system using explainable AI framework," Proc. Int. Conf. Cyber Security, 2021.
2. P. R. B. Houssel et al., "eX-NIDS: A framework for explainable network intrusion detection leveraging large language models," arXiv preprint arXiv:2507.16241, Jul. 2025.
3. A. Agarwal and M. J. Nene, "Incorporating AI incident reporting into telecommunications law and policy: Insights from India," Telematics Informatics Reports, vol. 14, p. 100132, 2024.
4. Y. Li et al., "IDS-Agent: An LLM agent for explainable intrusion detection in IoT networks," NeurIPS Workshop on Open-World Agents, 2024.
5. T. Sowmya and E. A. Mary Anita, "A comprehensive review of AI based intrusion detection system," Measurement: Sensors, vol. 28, 100827, 2023.
6. Z. T. Sworna, Z. Mousavi, and M. A. Babar, "NLP methods in host-based intrusion detection systems: A systematic review and future directions," J. Netw. Comput. Appl., vol. 220, p. 103761, 2023.

7. N. Albanbay et al., "Federated learning-based intrusion detection in IoT networks: Performance evaluation and data scaling study," J. Sensor Actuator Networks, vol. 14, no. 4, p. 78, 2024.
8. S. Lysenko et al., "The role of artificial intelligence in cybersecurity: Automation of protection and detection of threats," Econ. Aff., vol. 69, no. 1, pp. 43–51, 2024.
9. M. J. Olatunde, "Artificial intelligence for cybersecurity risk prioritization in complex digital ecosystems," Global Journal of Engineering and Technology Advances, vol. 20, no. 1, pp. 253–269, Jul. 2024.
10. P. Rafiey and A. Namadchian, "Mapping vulnerability description to MITRE ATT&CK framework by LLM," Adv. Artif. Intell. Mach. Learn., vol. 5, no. 3, pp. 4379–4396, Sep. 2025.
11. S. S. R. Chittoju et al., "Synergistic integration of blockchain and artificial intelligence for robust IoT and critical infrastructure security," Int. J. Adv. Res. Comput. Commun. Eng., vol. 14, no. 4, p. 778, Apr. 2025.
12. S. Krishnapriya and S. Singh, "A comprehensive survey on advanced persistent threat (APT) detection techniques," Comput. Mater. Continua, 2024.
13. N. Kumar and S. Sharma, "A hybrid modified deep learning architecture for intrusion detection system with optimal feature selection," Electronics, vol. 12, no. 19, 4050, 2023.
14. S. Mane and D. Rao, "Explaining network intrusion detection system using explainable AI framework," arXiv:2103.07110, 2021.
15. A. Abubakar Aliyu, J. Liu, and E. Gilliard, "A decentralized and self-adaptive intrusion detection approach using continuous learning and blockchain technology," J. Data Sci. Intell. Syst., 2024.
16. Y. Li et al., "IDS-Agent: An LLM agent for explainable intrusion detection in IoT networks," arXiv:2408.04817 (NeurIPS 2024 Workshop), 2024.
17. S. T. Sworna, Z. Mousavi, M. A. Babar, "NLP methods in host-based intrusion detection systems: A systematic review and future directions," IEEE Trans. Knowl. Data Eng., 2023 (preprint).
18. N. Albanbay, Y. Tursynbek, et al., "Federated learning-based intrusion detection in IoT networks: Performance evaluation and data scaling study," J. Sensor Actuator Networks, 2024.
19. M. Zhang et al., "Evaluating ML-based IDS with XAI," Frontiers in Comms. Netw., 2024.
20. X. Zhao et al., "Towards explainable network intrusion detection using LLMs," in Proc. IEEE ICC, 2024.