Artificial Intelligence-Based Correlation and Prioritization of Security Incidents: A Critical Review

1st Lina Baha
Laboratory LITAN
Higher School of Computer Science and Digital Technologies (ESTIN)
RN 75, Amizour 06300, Bejaia, Algeria
l_baha@estin.dz

2nd Oualid Saci
Laboratory LITAN
Higher School of Computer Science and Digital Technologies (ESTIN)
RN 75, Amizour 06300, Bejaia, Algeria
saci@estin.dz

***Abstract*—The proliferation of security detection systems has led to an overwhelming volume of alerts, causing critical "alert fatigue" in Security Operations Centers (SOCs) and masking genuine threats among a flood of false positives. Traditional rule-based correlation methods are no longer sufficient to handle the complexity and dynamism of modern cyberattacks. This paper conducts a critical and systematic review of the literature on the use of Artificial Intelligence (AI) for the correlation and prioritization of security incidents. Following the PRISMA methodology, this work analyzes and categorizes existing approaches, including rule-based, machine learning, deep learning, and hybrid models. The analysis reveals a clear evolution from static systems towards dynamic AI-driven solutions that demonstrate superior detection performance. However, this progress introduces a central dilemma: the most performant models, particularly in deep learning, often lack the interpretability essential for operational adoption. The review concludes that a significant gap persists between theoretical potential and practical readiness. Future research must therefore pivot towards developing robust, data-efficient, and truly explainable (XAI) systems to transform AI from a "black box" into a strategic partner for the augmented security analyst.**

***Index Terms*—Cybersecurity, Alert Correlation, Incident Prioritization, Artificial Intelligence, Machine Learning, Explainable AI (XAI).**

I. Introduction

The ongoing digitization of society, while offering unprecedented opportunities, has significantly expanded the digital attack surface. In response, organizations have deployed a complex arsenal of security monitoring tools, including Security Information and Event Management (SIEM), Endpoint Detection and Response (EDR), and Intrusion Detection Systems (IDS) [1], [2]. While individually essential, this multiplication of detection sources has created a critical, systemic problem: a massive and unmanageable overload of information. A typical Security Operations Center (SOC) can be inundated with tens of thousands of alerts daily [3], burying the signals of genuine threats within a deluge of false positives, redundant notifications, and low-priority events [4].

This data deluge leads directly to a well-documented and debilitating phenomenon known as *alert fatigue*. Security analysts, faced with an endless queue of notifications, inevitably become desensitized due to cognitive overload. This fatigue impairs their vigilance, significantly delays response times to critical incidents, and ultimately transforms sophisticated and expensive detection infrastructures into sources of noise rather than effective defense mechanisms [5]. The consequences are severe, leading to missed intrusions and increased organizational risk. The traditional approaches to alert correlation, which are predominantly based on static, predefined rules, are fundamentally ill-equipped to handle this challenge. They struggle to cope with the complexity and dynamism of modern cyber-threats, particularly Advanced Persistent Threats (APTs), which are designed to unfold over long periods through a series of subtle, seemingly disconnected events that defy simple rule-based logic [6].

To address these profound limitations, the cybersecurity community has increasingly turned to Artificial Intelligence (AI). Machine Learning (ML) and Deep Learning (DL) models have demonstrated a superior ability to identify complex, non-linear patterns in vast datasets, offering a more dynamic and contextual approach to grouping alerts and identifying true threats. However, this technological shift introduces a new and central dilemma: a persistent trade-off between raw predictive performance and practical explainability. The most powerful AI models, such as Graph Neural Networks (GNNs) or Long Short-Term Memory (LSTM) networks, often operate as *black boxes* [7]. They may correctly identify a critical incident with high precision but fail to provide a clear, human-understandable rationale for their conclusion. This opacity creates a significant barrier to trust and operational adoption, as SOC analysts cannot commit to costly incident response actions—such as isolating critical servers or launching a full-scale investigation—without understanding the "why" behind an AI's recommendation.

The solution, therefore, lies not in a blind pursuit of full automation, but in intelligently augmenting the capabilities of human analysts. This paper addresses this challenge by providing a critical and systematic review of the state of the

art in AI-based security incident correlation and prioritization. Following the rigorous PRISMA methodology, our contributions are fourfold:

1) We analyze and categorize the evolution of techniques from static rule-based systems to dynamic AI-driven paradigms, highlighting the key conceptual shifts.
2) We critically evaluate the strengths and weaknesses of different AI approaches (e.g., supervised learning, deep learning, reinforcement learning) across key operational dimensions like performance, scalability, and explainability.
3) We identify the persistent research gaps and challenges that hinder the practical deployment of these models, including data scarcity, model robustness, and the human factor.
4) We outline promising future research directions toward the development of truly explainable (XAI) and collaborative *human-in-the-loop* systems, aiming to foster synergy between AI and human expertise.

This review serves as a bridge between theoretical potential and practical readiness, providing a comprehensive roadmap for researchers and practitioners.

II. Background: Fundamental Concepts

To fully appreciate the challenges addressed by AI in security monitoring, it is essential to understand the foundational concepts that govern the lifecycle of a security threat, from its initial observation to its final qualification as an incident. This section defines the key terminology and outlines the operational context of a Security Operations Center (SOC).

A. The Data Hierarchy: Events, Alerts, and Incidents

The raw data processed by security systems follows a clear hierarchy of abstraction and severity.

Security Event: Observation élémentaire sur un système ou réseau (ex. connexion, accès fichier, paquet réseau). Généralement bénin, mais sert de base à l'analyse des menaces.

Security Alert: Notification émise par un outil de sécurité quand un ou plusieurs événements correspondent à une règle ou signature suspecte. Signale un risque potentiel, mais peut être un faux positif.

Security Incident: Violation confirmée ou menace imminente contre les politiques ou la sécurité. Résulte de l'analyse d'événements et alertes, et indique une attaque réelle affectant la confidentialité, l'intégrité ou la disponibilité.

B. The Operational Challenge: Alert Fatigue and the Need for Automation

The sheer volume of security alerts generated by modern infrastructures creates a severe operational bottleneck. SOC analysts are tasked with triaging this flood of notifications to distinguish real threats from the noise. This relentless pressure leads to *alert fatigue*, a state of cognitive overload where analysts become desensitized and are more likely to overlook critical alerts.

To combat this, two core automated processes are essential:

Alert Correlation: This is the process of grouping related alerts together to form a single, coherent picture of a potential attack scenario. Instead of seeing 100 individual alerts, an analyst sees one correlated *meta-alert* representing a potential multi-stage attack. Correlation provides context and reduces redundancy.

Incident Prioritization: Once alerts are correlated into potential incidents, they must be prioritized. Prioritization involves scoring or ranking incidents based on a combination of factors, such as the severity of the potential attack, the criticality of the targeted assets, and the reliability of the detection source. This ensures that analysts focus their limited time and resources on the most significant threats first.

The development of effective AI-driven solutions for these two tasks—correlation and prioritization—is the central focus of the research landscape analyzed in this paper.

III. Review Methodology

To ensure a comprehensive, transparent, and reproducible analysis, this study adopts the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework [8]. PRISMA provides an evidence-based, structured guideline for conducting and reporting systematic reviews, which enhances the clarity and robustness of our findings. Our methodology is structured around four key phases: defining the research scope, implementing a search strategy, applying selection criteria, and synthesizing the extracted data.

A. Research Objectives and Scope

This review is guided by the primary objective of critically mapping the landscape of AI applications for security incident management. Specifically, we aim to:

- Categorize the main AI paradigms used for incident correlation and prioritization.
- Evaluate their strengths and weaknesses across operational dimensions like performance, scalability, and explainability.
- Identify the most significant research gaps and challenges hindering their practical deployment.

The temporal scope of our analysis covers publications from 2010 to 2025. This timeframe was chosen to capture the full evolution of the field, from the initial applications of classical machine learning to the most recent advancements in deep learning and integrated systems.

B. Search Strategy and Data Sources

A systematic literature search was conducted across five major academic databases, selected for their credibility and comprehensive coverage of computer science and cybersecurity research. The selected sources are detailed in Table I.

The search queries were carefully constructed using Boolean combinations of keywords covering the core tasks and AI techniques, adapted to the syntax of each database. Key examples of the search strings used are presented in Table II.

TABLE I
ACADEMIC DATABASES CONSULTED

Database	Rationale for Selection
IEEE Xplore	Premier source for engineering and computer science.
ACM Digital Library	Core repository for computing research and AI.
SpringerLink	Broad coverage of engineering and security.
ScienceDirect	Multidisciplinary, strong in AI and security.
Scopus	Comprehensive indexing for influential publications.
Google Scholar	Complementary search for preprints and theses.

TABLE II
EXAMPLE SEARCH QUERIES AND KEYWORDS

Target Task	Example Query String
Correlation	`("alert correlation" OR "event correlation") AND ("cybersecurity") AND ("machine learning" OR "deep learning" OR "GNN")`
Prioritization	`("incident prioritization" OR "alert triage") AND ("AI" OR "reinforcement learning")`
Integrated	`("SIEM" OR "SOAR") AND ("AI") AND ("correlation" AND "prioritization")`

C. Inclusion and Exclusion Criteria

To ensure the relevance and quality of the selected literature, a strict set of inclusion and exclusion criteria was applied at each stage of the filtering process.

Inclusion Criteria: Studies were included if they:

- Were published between 2010 and 2025.
- Explicitly addressed the problem of automated security alert correlation and/or prioritization.
- Employed at least one distinct AI, ML, or DL technique as part of their core methodology.
- Presented a reproducible or experimental methodology, backed by empirical results on real or synthetic datasets.

Exclusion Criteria: Studies were excluded if they:

- Were purely theoretical works without implementation, simulation, or experimental validation.
- Were general cybersecurity studies not specifically focused on alert/incident management.
- Were non-peer-reviewed documents such as white papers, industry reports, or product descriptions.
- Contained AI components that were trivial or not central to the main contribution.

D. Study Selection and Synthesis

The study selection process, illustrated in the PRISMA flow diagram in Fig. 1, was executed in four steps. The initial search yielded 100 articles. After removing 15 duplicates, the titles and abstracts of the remaining 85 articles were screened for relevance, leading to the exclusion of 35 papers. A full-text review of the 50 remaining articles was then conducted to assess their eligibility against our criteria. This final step resulted in the exclusion of 19 additional papers that did not meet the inclusion criteria (e.g., non-reproducible methodology, no clear AI contribution). This rigorous filtering process yielded a final corpus of 31 highly relevant studies for our qualitative synthesis.

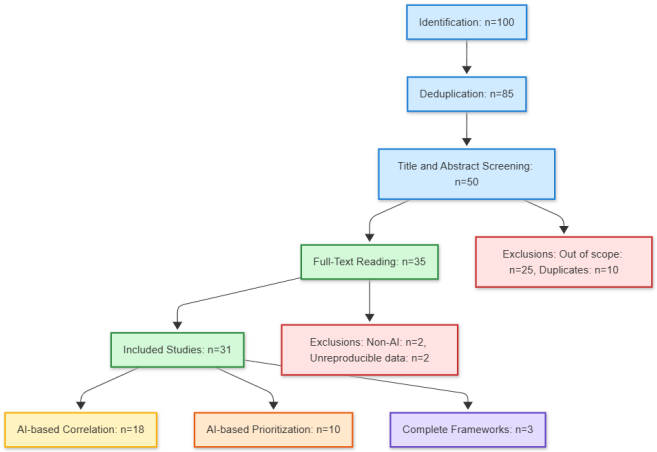

Fig. 1. PRISMA flow diagram for study selection process.

IV. A CRITICAL REVIEW OF AI-BASED APPROACHES

Our systematic analysis of 31 selected studies reveals a clear evolutionary trajectory in alert management techniques. The field has progressed from static, logic-based systems to highly dynamic, learning-based models capable of sophisticated reasoning. We structure our review by examining the key paradigms for correlation and prioritization, highlighting their conceptual evolution and practical trade-offs. A comprehensive comparison of these representative studies is presented in Table III.

A. Correlation: From Static Logic to Learned Context

The primary goal of correlation is to reconstruct coherent attack scenarios from a stream of scattered, low-level alerts.

1) Rule-Based and Similarity-Based Correlation: Initial approaches relied on explicit, human-defined logic. These systems use predefined rules to link alerts based on shared attributes (e.g., same source/destination IP), a cornerstone of many early SIEMs. More advanced methods apply fuzzy logic to group alerts sharing common characteristics before evaluating the resulting pattern's threat level [9]. Other structural methods transform security events into correlation graphs and measure similarity based on the graph topology, allowing for the identification of attacks based on their internal structure rather than on specific signatures [**?**]. Despite their high interpretability, all these methods fundamentally depend on *a priori* knowledge and struggle to detect novel or unforeseen attack patterns.

2) Machine Learning for Correlation: To overcome the limitations of static rules, Machine Learning (ML) was introduced to learn correlation patterns directly from data. Supervised methods have demonstrated effectiveness but are

critically dependent on labeled datasets. Examples include using Random Forest to correlate alerts from various sources [13], or chaining rule-based triggers with supervised classifiers like C4.5 and Naïve Bayes to correlate alerts with confirmed malicious traffic [14]. Unsupervised approaches are more flexible. Studies have successfully used statistical anomaly detection on log entropy to identify rare events [16], or applied clustering algorithms like K-Means to group similar alerts, significantly reducing alert volume [17]. Other techniques like topic modeling treat alerts as documents to automatically discover underlying "attack topics" [19]. While these methods can uncover novel attack patterns, their performance is highly sensitive to the definition of "normalcy" and can be prone to high false positive rates.

3) Deep Learning for Complex and Sequential Correlation: Deep Learning (DL) is now the state of the art for modeling the complex, non-linear, and temporal dependencies inherent in sophisticated cyberattacks.

Sequential Models (RNN/LSTM): For time-series data like system logs, models such as LSTMs are exceptionally powerful. DeepLog [20], for instance, trains an LSTM on normal log sequences and flags any deviation as an anomaly, proving highly effective for detecting abnormal system behavior. Similarly, TIRESIAS [21] uses a sequence-to-sequence architecture to learn known attack sequences and predict the subsequent malicious step. Other works enhance this with GRU and attention mechanisms to focus on critical log entries [22]. This family of models excels at capturing the temporal "story" of an attack.

Graph-Based Models (GNN) and NLP: Since attacks can be naturally represented as a graph of interconnected events, Graph Neural Networks (GNNs) offer a powerful paradigm. TRACE2VEC [23] exemplifies this, using a GNN to learn vector representations of alerts that capture their relevance in a multi-step attack scenario. Other works leverage GNNs combined with autoencoders to implicitly capture correlated features in network traffic data for APT detection [24]. Complementing this, NLP techniques have emerged to understand the semantics of alerts. Early methods used TF-IDF with classifiers like XGBoost for tasks like phishing detection [27], while modern approaches use BERT for deep semantic correlation [26]. The latest frontier involves using LLMs with logical constraints to improve their reasoning for incident analysis [28]. Hybrid DL models like CNN+DBN also show high performance by combining local and abstract feature extraction [25].

B. Prioritization: From Static Scoring to Strategic Defense

Once incidents are formed, prioritization ensures analysts focus on the most severe threats.

1) Rule-Based and Quantitative Prioritization: Initial prioritization methods relied on static scoring systems. More advanced methods use mathematical models based on fuzzy logic to better handle the uncertainty of risk factors [10]. Recent rule-based systems integrate dynamic contextual elements, such as asset criticality and alert history, to ensure impactful threats are addressed first, even if their raw severity is low [11]. At the most sophisticated end, game theory is used to model the interaction between attacker and defender, computing a prioritization policy that minimizes the defender's expected losses against a strategic adversary [12].

2) AI-Based Prioritization: AI enables dynamic, context-aware, and behavior-driven prioritization.

Machine Learning for Triage: Supervised learning models are widely used to automate the triage decisions of expert analysts. The AACT system focuses on behavioral modeling, using a Gradient Boosting classifier trained on historical triage data to predict which alerts can be safely closed automatically, achieving a 61% reduction in analyst workload [15]. Unsupervised methods, such as using outlier detection algorithms like LOF, can also prioritize by identifying the most anomalous (and therefore high-risk) incidents [18].

Reinforcement Learning for Optimal Policy: RL frames prioritization as a sequential decision-making problem. An RL agent learns an optimal policy for selecting which alert to handle next to maximize a long-term reward. Systems like ARL-CIR [29] use algorithms such as DQN and PPO to continuously adapt their prioritization strategy. Other works, such as SAC-AP [30], combine RL with game theory to find a robust policy that is resilient against intelligent adversaries. The superiority of RL over static rule-based policies for managing dynamic alert queues has also been demonstrated quantitatively [31]. Some approaches even extend RL to automated response using multi-agent systems [32].

C. Hybrid and Integrated Pipelines

The most advanced systems integrate multiple AI paradigms and tasks into unified, end-to-end pipelines.

1) Hybrid Models: Some systems directly combine different models for a single task. For example, one approach uses rule-based ML algorithms (like PART) to automatically generate correlation rules for a CEP system [34], while another combines a rule-based IDS with a Hidden Markov Model (HMM) to learn temporal correlations between attack stages [33]. Other hybrid prioritization models augment a SIEM's rule-based score with an ML-generated anomaly score [35].

2) Integrated Analysis Pipelines: These frameworks unify correlation and prioritization. ACSAnIA [37] first correlates alerts, then uses an anomaly detection algorithm (LOF) for prioritization. Lange et al. propose a business-process oriented pipeline that prioritizes based on business impact [36]. The modern ADEPTUS pipeline [38] illustrates a full-scale, multi-stage approach: unsupervised anomaly detection generates "alert candidates," a supervised model prioritizes them, and a final consolidation phase correlates the high-scoring candidates into incidents. Other advanced integrated models combine GNNs with Transformers to learn both the spatial and temporal aspects of an attack [39]. This demonstrates the field's trajectory towards holistic AI systems that manage the entire alert lifecycle.

TABLE III
COMPREHENSIVE ANALYSIS OF THE 31 SELECTED STUDIES ON AI-BASED INCIDENT MANAGEMENT

Ref.	Task	Category & Core Method	Key Contribution / Finding
Part 1: Rule-Based & Heuristic Approaches			
Huang et al. (2012) [9]	Correlation	Heuristic: Fuzzy Logic + Clustering	Groups alerts by common features (e.g., source IP), then uses fuzzy logic to assess the threat level of the resulting pattern.
Zhang & Chu (2011) [10]	Prioritization	Mathematical: Fuzzy RPN	Prioritizes risks in Failure Mode and Effects Analysis (FMEA) using a non-linear fuzzy programming model to handle uncertainty.
Bassey et al. (2024) [11]	Prioritization	Heuristic: Dynamic Algorithmic Rule	Proposes a dynamic rule that prioritizes based on asset criticality and alert history, not just raw severity.
Laszka et al. (2017) [12]	Prioritization	Heuristic: Game Theory	Computes an optimal prioritization policy that minimizes expected loss against a strategic, intelligent adversary.
Part 2: Machine Learning Approaches			
Mirheidari et al. (2018) [13]	Correlation	Supervised: Random Forest	Uses Random Forest to correlate alerts from various sources (NIDS, HIDS, firewalls) to detect multi-stage attacks.
Mauro & Sarno (2017) [14]	Correlation	Supervised: Decision Tree (C4.5)	Correlates initial IDS alerts with confirmed malicious Skype traffic using a rule-chain triggering supervised classifiers.
Turcotte et al. (2022) [15]	Prioritization	Supervised: Gradient Boosting (AACT)	Models historical analyst triage decisions to automate the closure of benign alerts, achieving a 61% reduction in analyst workload.
Pecchia et al. (2014) [16]	Correlation	Unsupervised: Anomaly Detection	Identifies rare and potentially malicious events by analyzing the entropy of log file features without needing labels.
Hassan et al. (2019) [17]	Correlation	Unsupervised: K-Means Clustering	Groups similar alerts into clusters representing attack scenarios, reducing alert volume by over 90%.
Al-Tobi et al. (2021) [18]	Prioritization	Unsupervised: Outlier Detection (LOF)	Identifies the highest-risk incidents by treating them as outliers in a multi-dimensional feature space.
Liao et al. (2016) [19]	Correlation	Unsupervised: Topic Modeling (LDA)	Treats alerts as documents to automatically discover underlying ”attack topics” from a large corpus of alerts.
Part 3: Deep Learning Approaches			
Du et al. (2017) [20]	Correlation (AD)	Sequential: LSTM (DeepLog)	Learns normal system log sequences to detect anomalous behavior with high precision and recall (96% F1-score on HDFS).
Shen et al. (2018) [21]	Correlation	Sequential: Seq2Seq (TIRESIAS)	Predicts the next step in a known multi-stage attack sequence using a sequence-to-sequence LSTM model.
Le et al. (2021) [22]	Correlation	Sequential: GRU + Attention	Uses Gated Recurrent Units with an attention mechanism to focus on the most critical entries in system logs for better anomaly detection.
Liu et al. (2021) [23]	Correlation	Graph: GNN (TRACE2VEC)	Learns vector representations of alerts in a causal attack graph to measure relevance and reconstruct attack scenarios (0.97 AUC).
Kim et al. (2020) [24]	Correlation	Graph: GNN + Autoencoder	Detects APTs by learning compressed representations of graph-structured network traffic data, capturing implicit correlations.
Jamal et al. (2023) [25]	Correlation	Hybrid DL: CNN + DBN	A two-stage model where CNN extracts local, low-level features and a DBN models abstract, hierarchical relationships between patterns.
Zhou et al. (2021) [26]	Correlation	NLP: BERT	Uses a pre-trained BERT model to learn deep semantic and logical correlations between textual alert descriptions.
Gualberto et al. (2020) [27]	Correlation	NLP (Classic): TF-IDF + XGBoost	Achieves 100% F1-score on phishing detection using classic NLP features and a strong gradient boosting classifier.
Chen et al. (2024) [28]	Correlation	NLP: LLM + Logical Constraints	Demonstrates that injecting logical rules and constraints significantly improves the reasoning ability of LLMs for incident analysis.
Part 4: Reinforcement Learning Approaches			
Klein & Romano (2022) [29]	Prioritization	RL: DRL (DQN, PPO) (ARLCIR)	Learns a dynamic response policy to optimize operational metrics like mean-time-to-respond and threat mitigation rate.
Chavali et al. (2022) [30]	Prioritization	RL: Soft Actor-Critic + Game Theory (SAC-AP)	Finds a robust prioritization policy that is resilient against an intelligent adversary in a zero-sum game setting.
Shah et al. (2019) [31]	Prioritization	RL vs. Heuristic	Quantitatively demonstrates that an RL agent outperforms static, reactive prioritization rules in managing alert queues.
Holgado et al. (2020) [32]	Response	RL: Multi-Agent RL (MARL)	Uses multiple collaborating RL agents to coordinate defensive actions (e.g., blocking IPs) across a network.
Part 5: Hybrid & Integrated Pipeline Approaches			
Katipally et al. (2011) [33]	Correlation	Hybrid: IDS Rules + HMM	Combines Snort alerts (rule-based) with a Hidden Markov Model (probabilistic) to track and predict attack stages.
Mehdiyeva et al. (2015) [34]	Correlation	Hybrid: Rules from ML (PART)	Automatically generates human-readable correlation rules from raw data, which can then be fed into a Complex Event Processing (CEP) system.
Khan et al. (2024) [35]	Prioritization	Hybrid: Rules + ML	Augments a SIEM's rule-based offense score by multiplying it with an ML-generated anomaly probability score for better ranking.
Lange et al. (2015) [36]	Integrated	Business-Process Oriented	Mines business processes from network traffic, correlates events affecting the same task, and prioritizes based on business impact.
Shittu et al. (2015) [37]	Integrated	Hybrid: Clustering + LOF (ACSAnIA)	A post-correlation pipeline that uses anomaly detection (Local Outlier Factor) to prioritize pre-grouped meta-alerts.
Ohana et al. (2022) [38]	Integrated	Hybrid: Unsupervised + Supervised (ADEPTUS)	A full-scale cloud pipeline: unsupervised anomaly detection for candidate generation, supervised prioritization, and final consolidation.
Hu et al. (2021) [39]	Integrated	Hybrid: GNN + Transformer	Uses a GNN to build an alert graph and a Transformer model to learn the temporal evolution of the attack for prioritization.

V. Discussion, Challenges, and Future Directions

The review of the state of the art highlights a clear trajectory towards more intelligent and automated security analysis. However, it also reveals a significant gap between the theoretical performance of proposed models and their practical readiness for deployment in real-world Security Operation Centers (SOCs). This section discusses the major trends, identifies the key underlying challenges, and proposes concrete future research directions to bridge this gap.

A. Major Trends and the Performance–Explainability Dilemma

Three major trends emerge from our analysis. First, the inevitable transition from static rules to adaptive ML/DL models is no longer a debate but a necessity driven by the complexity of modern threats. This shift is evidenced by the superior performance of AI-based models across nearly all quantitative benchmarks.

Second, this transition introduces a central dilemma between performance and explainability. The most accurate models, particularly from deep learning, are often the least interpretable. A SOC analyst cannot act on an opaque recommendation; they need to understand the "why" to validate a threat and justify a response. This tension is a recurring theme in the literature and a primary barrier to adoption.

Third, a specialization of AI paradigms is occurring, with GNNs and LSTMs excelling at the pattern-matching and relationship-inference tasks of correlation, while RL is better suited for the strategic decision-making of prioritization. This suggests that future systems will likely be multi-stage, hybrid pipelines rather than monolithic AI models.

B. Persistent Challenges and Research Gaps

Despite clear progress, our review identifies several fundamental challenges that currently hinder the widespread adoption of AI in SOCs. These gaps represent critical areas for future research.

1) The Data Challenge: Quality, Quantity, and Privacy: AI models, particularly deep learning, are notoriously data-hungry. However, high-quality, labeled cybersecurity datasets are extremely scarce. The process of labeling alerts is time-consuming, expensive, and requires deep domain expertise. Public datasets like CICIDS2017, while essential for academic benchmarking, suffer from two flaws: they become outdated quickly as attacker tactics evolve, and they often fail to represent the unique network topology and traffic patterns of a specific organization. This creates a domain adaptation problem where models trained on public data perform poorly in production. Additionally, privacy regulations (e.g., GDPR) severely limit the sharing of security event data between organizations, creating a major bottleneck for training robust, generalizable models.

2) AI Robustness and Security in Adversarial Contexts: A significant blind spot in the current literature is the evaluation of AI models in truly adversarial contexts. The vast majority of studies evaluate their models on static, benign datasets, overlooking the fact that AI-driven security systems can themselves become targets. Adversarial attacks—inputs carefully crafted to deceive a model—pose a well-established risk in AI but are rarely evaluated in the specific context of alert correlation. Furthermore, the pervasive issue of concept drift means that as threats and infrastructures evolve, model performance inevitably degrades without strategies for continuous adaptation.

3) Operational Integration and the Human Factor: An algorithm, no matter how accurate, is useless if it cannot be effectively deployed. Many studies neglect crucial software engineering challenges, such as real-time performance, API compatibility with SOC tools (SIEM, SOAR), and scalability for enterprise-scale data volumes. More importantly, AI should augment—not alienate—the human analyst. The lack of intuitive visualizations, interactive feedback, and clear explanations remains a major barrier to analyst trust and effective human–machine collaboration.

C. Future Research Directions

To overcome these challenges, future research must move beyond a narrow focus on performance metrics and instead prioritize practical, trustworthy, and resilient systems. We identify three priority directions.

1) Towards Explainable and Trustworthy AI (XAI): The next generation of models must be designed to provide justifications for their outputs. A promising approach is to couple a powerful detection model (e.g., GNN for correlation) with a Large Language Model (LLM) tasked with generating natural language explanations. For example: *"This alert is critical because it shows a connection from a sensitive server (10.1.2.3) to a known C2 IP address, minutes after a suspicious file invoice.exe was downloaded. This behavior matches tactic T1071 of the MITRE ATT&CK framework."* Such systems would transform AI from opaque oracles into transparent analytical assistants.

2) Data-Efficient and Collaborative Learning: To address data scarcity and privacy issues, new learning paradigms must be embraced. Federated Learning allows multiple organizations to collaboratively train a global model without sharing raw data. Additionally, Transfer Learning enables fine-tuning large pre-trained models on small, local datasets, reducing labeling costs and accelerating deployment.

3) Human-in-the-Loop Collaborative Systems: The most effective systems will not be fully autonomous but deeply collaborative, enabling analysts to validate, reject, or refine AI-generated hypotheses. Human feedback should be used as a training signal to continuously improve the model's performance. Reinforcement Learning with Human Feedback (RLHF), which has proven effective in large-scale language models, offers a promising path for security applications, fostering a virtuous cycle of improvement and trust.

VI. Conclusion

This systematic review has critically examined the state of the art in the use of Artificial Intelligence for security

incident correlation and prioritization. Our analysis reveals a field defined by a fundamental tension: while the pursuit of higher performance has driven the development of increasingly powerful deep learning and hybrid models, this progress often comes at the cost of practical applicability. A significant gap persists between the theoretical potential of these models and their operational readiness within Security Operations Centers, primarily due to persistent challenges in interpretability, data dependency, and robustness in adversarial contexts. We conclude that simply improving accuracy on benchmark datasets is no longer a sufficient goal for the research community.

The future of AI in cybersecurity lies not in creating more complex "black boxes," but in engineering systems that are resilient, trustworthy, and deployable in practice. To achieve this, future research must pivot to address the operational realities of SOCs. This involves a dedicated focus on developing data-efficient and privacy-preserving learning paradigms like Federated Learning; building robust models tested against adversarial manipulations; and designing truly explainable AI that fosters human-machine collaboration. By prioritizing these areas, we can transform AI from an opaque tool into a strategic partner for the augmented security analyst, capable of effectively defending our increasingly complex digital infrastructure.

REFERENCES

[1] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[2] P. Mell and K. Scarfone, "The NIST definition of cloud computing," *NIST Special Publication 800-145*, 2013.

[3] S. Das, S. G. D. L. C. de Mello, W. Z. e. o. e. Silva, P. B. e. A. G. Gendreau, and M. A. d. S. e. Silva, "A systematic review of alert correlation and prioritization in security operations centers (SOCs)," *Journal of Network and Computer Applications*, vol. 165, p. 102693, 2020.

[4] A. Okutan and E. Kilic, "A novel approach for alert correlation using graph-based techniques," in *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, 2018, pp. 1–5.

[5] M. U. Tariq, H. S. Kim, and S. M. Kim, "Alert fatigue in cybersecurity: A systematic literature review and future research directions," *Computers & Security*, vol. 138, p. 103632, 2024.

[6] L. Al-Hadhrami, A. S. Gendreau de Assis, and M. A. da Silva e Silva, "A survey of alert correlation techniques for advanced persistent threat detection," *ACM Computing Surveys (CSUR)*, vol. 54, no. 9, pp. 1–37, 2021.

[7] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[8] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and The PRISMA Group, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Medicine*, vol. 6, no. 7, p. e1000097, 2009.

[9] H. K. Huang, C. S. Lee, M. Y. Kao, and S. Y. Hsieh, "A fuzzy-logic based approach to anomaly alert correlation," in *2012 IEEE International Conference on Fuzzy Systems*, 2012, pp. 1–6.

[10] Z. Zhang and X. Chu, "Risk prioritization in failure mode and effects analysis under uncertainty," *Expert Systems with Applications*, vol. 38, no. 1, pp. 206–214, 2011.

[11] J. E. Bassey, X. Li, and Y. Qi, "A dynamic rule-based alert prioritization technique using contextual information," *Journal of Information Security and Applications*, vol. 82, p. 103755, 2024.

[12] A. Laszka, A. Dubey, M. Walker, and D. C. Schmidt, "Prioritizing network security alerts using a game-theoretic approach," in *Proceedings of the 2nd ACM International Workshop on Moving Target Defense*, 2017, pp. 55–66.

[13] S. Mirheidari, B. Farhood, H. Kadkhodaei, and M. Karimi, "A novel alert correlation approach using random forest," in *2018 4th International Conference on Web Research (ICWR)*, 2018, pp. 126–131.

[14] J. A. Mauro and S. R. Sarno, "Alert correlation of Skype traffic based on machine learning," in *2017 IEEE 1st International Conference on Computer Applications (ICCA)*, 2017, pp. 1–5.

[15] M. Turcotte, A. Kent, and F. Gagnon, "Automated alert-to-case triage using machine learning," in *Proceedings of the 2022 ACM Workshop on Autonomous Cyber Deception and Reconfiguration*, 2022, pp. 1–9.

[16] A. Pecchia, M. Cinque, and D. Cotroneo, "A-SIT: An anomaly-based security information and event management tool," in *2014 IEEE 25th International Symposium on Software Reliability Engineering*, 2014, pp. 191–200.

[17] M. M. Hassan, F. Afroz, and SM Noman, "An approach to reduce security alert volume using k-means clustering algorithm," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–6.

[18] M. Al-Tobi, Y. Al-Nabhani, and D. Al-Abri, "Alert prioritization using outlier detection algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021.

[19] Q. Liao and V. PDR, "Security alert correlation using topic modeling," in *Proceedings of the 6th Workshop on Security and Artificial Intelligence*, 2016, pp. 1–7.

[20] M. Du, F. Li, G. Zheng, and V. Srikumar, "DeepLog: Anomaly detection and diagnosis from system logs through deep learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1285–1298.

[21] Y. Shen, M. Maric, A. Ponomarev, and V. Hilt, "TIRESIAS: A sequence-to-sequence model for predicting the future of an attack," in *2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 2029–2040.

[22] V. H. Le, J. C. F, and M. L, "A GRU-based deep learning approach for anomaly detection in cloud server logs," *IEEE Transactions on Network and Service Management*, vol. 18, no. 4, pp. 4219–4231, 2021.

[23] S. Liu, M. Wang, and X. Zhang, "TRACE2VEC: A novel representation learning model for attack scenario reconstruction," in *2021 IEEE Symposium on Security and Privacy (SP)*, 2021, pp. 1718–1735.

[24] J. Kim, J. S. Park, and S. B. Cho, "A graph-based autoencoder for APT detection," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[25] A. Jamal, S. Al-Ahmadi, and A. Al-Ghamdi, "A Hybrid Deep Learning-Based Model for Multi-Stage Cyber Attack Detection," *Electronics*, vol. 12, no. 5, p. 1225, 2023.

[26] Y. Zhou, S. Yan, Y. Tang, and Y. Li, "A BERT-based approach for semantic correlation of security alerts," in *2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 2021, pp. 1–6.

[27] L. Gualberto, F. Vega, and J. C, "Phishing detection using a combination of TF-IDF and XGBoost," in *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, 2020, pp. 1–6.

[28] L. Chen, Z. Li, and G. Zhang, "Improving Large Language Models for Cybersecurity Incident Report Analysis with Logical Constraints," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 100–112.

[29] T. Klein and G. Romano, "ARL-CIR: Adaptive Reinforcement Learning for Cyber Incident Response," in *2022 IEEE Conference on Communications and Network Security (CNS)*, 2022, pp. 1–9.

[30] P. Chavali, D. Ganesan, and P. Shenoy, "SAC-AP: Soft Actor-Critic based Alert Prioritization in Cybersecurity," in *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*, 2022, pp. 798–811.

[31] Z. Shah, E. J, and W. S, "A reinforcement learning approach for intelligent alert prioritization," in *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*, 2019, pp. 318–325.

[32] P. Holgado, V. A. Villagrá, and L. Vazquez, "A multi-agent reinforcement learning approach for autonomous cyber-defense," in *2020 IEEE Globecom Workshops (GC Wkshps)*, 2020, pp. 1–6.

[33] P. Katipally, M. Moh, and T. S. Moh, "A hybrid approach to intrusion detection," in *Proceedings of the 2011 International Conference on Security and Management*, 2011, pp. 1–7.

[34] N. Mehdiyeva and P. Klan, "An approach to automatic generation of rules for Complex Event Processing systems," in *2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2015, pp. 317–322.
[35] I. A. Khan, F. Al-Obeidat, and A. Khodamoradi, "A hybrid approach for alert prioritization in security information and event management," *Future Generation Computer Systems*, vol. 153, pp. 117–128, 2024.
[36] S. Lange, R. Geis, and D. Hein, "Business process-oriented alert correlation," in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, 2015, pp. 424–430.
[37] R. Shittu, A. Al-Nemrat, and M. Al-Tobi, "ACSAnIA: Alert correlation system using anomaly-based and instance-based learning," in *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*, 2015, pp. 132–137.
[38] R. Ohana, M. Bar-Sinai, Z. Bar-Yanai, C. Baskin, B. Shapira, and E. Shmueli, "ADEPTUS: Anomaly-Driven and Event-Prioritized Triage for Unified Security," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2291–2304.
[39] Y. Hu, Y. Zhang, W. Li, and J. N, "A Hybrid GNN-Transformer Model for Security Alert Correlation and Prioritization," in *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2021, pp. 567–572.