iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VIII **Month of publication:** August 2026

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

Artificial Intelligence for Real-Time Cyber Threat Classification and Emerging Threat Detection: A Structured Review of Methods, Datasets, Challenges, and Research Directions

Jaswanth Syam Sundar Garugu

Abstract: ***The swift adoption of cloud computing, the Internet of Things (IoT), fifth-generation (5G) communication, and edge computing has extensively broadened the cyber-attack surface, facilitating the emergence of threats with increased scale, speed, and complexity. Traditional detection techniques, which rely on signatures, rules, and statistical analysis, continue to be effective for identifying known attacks; however, their capacity to detect zero-day exploits, polymorphic malware, and advanced persistent threats is limited. Consequently, there has been a rise in the implementation of artificial intelligence (AI) for adaptive and real-time cyber threat analysis. This review evaluates the evolution of AI-driven methodologies for real-time cyber threat classification and the identification of new threats, drawing from over fifty peer-reviewed studies identified through a structured search of leading scholarly databases. Instead of treating the studies in isolation, the literature is synthesized based on detection objectives, computational techniques, datasets, evaluation metrics, deployment environments, and acknowledged limitations. The review juxtaposes conventional detection methods with machine learning, deep learning, explainable AI (XAI), reinforcement learning, federated learning, graph neural networks (GNNs), large language models (LLMs), and generative AI. It places particular emphasis on frequently utilized cybersecurity datasets and evaluation practices, as well as ongoing challenges related to class imbalance, adversarial manipulation, computational overhead, model interpretability, privacy concerns, and inadequate validation in real-world scenarios. The reviewed literature indicates that AI-based methodologies often demonstrate superior detection capabilities for intricate and previously unseen attack patterns compared to traditional methods; however, direct performance comparisons are complicated due to discrepancies in datasets, experimental designs, and evaluation protocols. Moreover, a limited number of proposed models have been evaluated under authentic operational circumstances. In light of these observations, the review identifies significant research gaps and outlines prospective paths for developing explainable, privacy-conscious, computationally efficient, and readily deployable AI-enabled cyber-defence systems.***

Keywords: ***Artificial Intelligence; Cyber Threat Detection; Real-Time Threat Classification; Machine Learning; Deep Learning; Emerging Threats; Explainable AI; Federated Learning; Graph Neural Networks; Large Language Models; Generative AI; Cybersecurity.***

I. INTRODUCTION

The advancement of digital technologies has transformed communication, information storage, and service delivery for individuals, businesses, and governments. Technologies such as cloud computing, the Internet of Things (IoT), fifth-generation (5G) networks, edge computing, and digital financial systems have created a highly interconnected landscape. While these innovations enhance efficiency, scalability, and accessibility, they also expand the range of potential attack vectors. Consequently, cyber threats have become more frequent and sophisticated, increasing the risk of financial, operational, and reputational harm across various sectors [2], [20].

Recent attacks have evolved from singular malicious actions to coordinated operations that involve various tactics such as malware, ransomware, phishing, botnets, insider threats, distributed denial-of-service (DDoS) attacks, advanced persistent threats (APTs), and zero-day exploits. Attackers are increasingly employing evasion techniques, polymorphic malware, encrypted communication channels, and methods enhanced by artificial intelligence, which presents significant challenges to traditional defence mechanisms that need to detect and respond immediately [24], [25].

Conventional detection methods primarily depend on signature-based, rule-based, and statistical approaches.

While these techniques effectively identify known threats with a high degree of accuracy and minimal false positives, they face challenges when addressing new patterns and the rapid evolution of threats. Signature-based systems are unable to detect attacks that do not have a pre-existing signature, and rule-based systems necessitate ongoing expert maintenance, which becomes increasingly difficult to manage in dynamic network environments. [7], [9], [22], [26]. The described constraints account for the transition toward intelligent, adaptive, and automated defence mechanisms.

Artificial intelligence (AI) facilitates this transition by processing extensive amounts of security data, identifying underlying patterns, recognizing anomalies, categorizing malicious activities, and aiding in timely decision-making. Technologies such as machine learning, deep learning, reinforcement learning, explainable AI (XAI), and graph-based learning have enhanced the accuracy, precision, and scalability of threat detection, thereby assisting organizations in minimizing false positives and fortifying security operations.[1], [10], [24]. The growing adoption of AI technologies has led to the development of various frameworks utilizing supervised learning, unsupervised learning, hybrid models, deep neural networks, transformer architectures, and explainable AI methods. However, variances in datasets, evaluation metrics, feature engineering techniques, and deployment environments hinder effective comparison of these frameworks and the identification of the optimal solution. Challenges such as class imbalance, adversarial attacks, privacy concerns, computational expenses, limited interpretability, and scalability issues persist, complicating practical deployment. [10], [20], [36].

A. Objectives and Research Questions

This review analyzes AI-enhanced real-time cyber threat classification and the detection of emerging threats through a comprehensive examination of recent studies. It discusses both traditional and AI-based detection methods, surveys relevant benchmark datasets and evaluation metrics, compares various research contributions, identifies existing challenges and gaps, and proposes future research directions for developing intelligent, adaptive, and reliable cyber defence systems. The review is structured around four research questions:

RQ1. How do AI-driven approaches enhance the classification of real-time cyber threats compared to conventional detection methods, and what techniques are leading in current practice?

RQ2. What benchmark datasets and evaluation metrics are used to support the reported results, and in what ways do their limitations impact the comparability across different studies?

RQ3. What persistent methodological limitations, challenges, and research gaps hinder AI-driven cyber threat detection?

RQ4. Which research directions show the most potential for the development of explainable, privacy-preserving, adaptive, and deployable cyber-defence systems?

B. Contributions

This review presents the following contributions:

•The text synthesizes AI-driven cyber threat detection by integrating traditional methods, classical machine learning, deep learning, and emerging techniques into a single taxonomy, rather than presenting them as isolated descriptions.

•It explores advanced methodologies, including explainable AI (XAI), federated learning (FL), graph neural networks (GNNs), large language models (LLMs), and generative AI, while linking each to specific cybersecurity applications [32], [34], [35], [43], [50].

•The analysis compares existing studies thematically based on technique, dataset, evaluation, and practical limitations, consolidating points of consensus, tension, and unresolved issues within the literature.

•It identifies a clear, evidence-based research gap and delineates priorities for developing accurate, scalable, explainable, and reliable AI-driven cybersecurity solutions.

C. Structure of the Review

Section 2 presents the methodology employed for the review. Section 3 examines the current cyber threat landscape. Section 4 assesses traditional detection methods. Section 5 elaborates on classical machine learning and deep learning techniques for threat classification and provides a taxonomy of these approaches. Section 6 investigates emerging methodologies. Section 7 reviews benchmark datasets, while Section 8 evaluates assessment metrics. Section 9 offers a comparative synthesis. Sections 10 and 11 focus on challenges and highlight research gaps. Section 12 outlines future directions, and Section 13 summarizes the findings.

II. REVIEW METHOD

This review follows a structured, reproducible procedure to analyse recent work on AI-enhanced real-time cyber threat classification and emerging threat detection. The aim is to understand existing research, record technical progress, assess limitations, identify gaps, and set out research opportunities. The procedure is presented as a structured literature review rather than a registered systematic review or meta-analysis, because verified screening counts and a pre-registered protocol were not maintained. Figure 1 summarises the process.

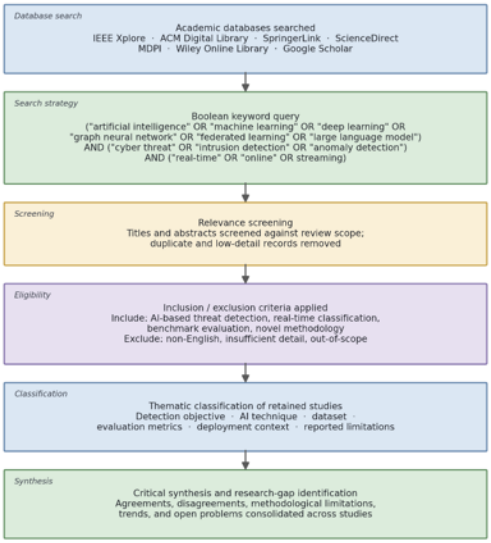


Fig. 1. Structured review process, from database search to critical synthesis.

A. *Sources and Search Strategy*

The search encompassed peer-reviewed journal articles, conference proceedings, book chapters, and esteemed scholarly publications in the domains of cybersecurity and artificial intelligence. Relevant materials were identified through various platforms, including IEEE Xplore, the ACM Digital Library, SpringerLink, ScienceDirect, MDPI, Wiley Online Library, and Google Scholar. To ensure reproducibility, primary concepts were integrated using Boolean logic, represented as follows: ("artificial intelligence" OR "machine learning" OR "deep learning" OR "graph neural network"). The methodology emphasized identifying studies related to federated learning, large language models, cyber threats, intrusion detection, anomaly detection, cybersecurity, and real-time or online environments. Publications from 2021 to 2026 were prioritized to reflect recent advancements in the field, with earlier research included when it provided significant theoretical frameworks, methodological insights, or key cybersecurity datasets relevant to the review.

B. *Inclusion and Exclusion Criteria*

Eligibility criteria for studies included a focus on AI-based cyber threat detection, real-time threat classification, intelligent intrusion detection, anomaly detection, or novel AI applications in the field of cybersecurity. Additionally, it was necessary for studies to present a clear methodology, provide a thorough evaluation, or utilize an established benchmark. Studies were excluded if they lacked technical detail, replicated previous findings, were published in a non-English language, or did not directly relate to AI-driven cybersecurity.

C. *Analysis and Synthesis*

The studies reviewed were categorized based on their research objectives, AI methodologies, datasets utilized, evaluation metrics employed, application fields, and noted strengths and weaknesses. A comparative analysis of each study was conducted with regard to detection goals, techniques, dataset attributes, evaluation criteria, deployment contexts, and acknowledged limitations. Special emphasis was placed on aspects such as model performance, computational efficiency, scalability, interpretability, resistance to adversarial attacks, and preparedness for real-time deployment[10], [37], [38].The synthesis highlighted emerging trends, persistent challenges, and research gaps, forming the foundation for the subsequent discussion [6], [10].

III. THE CYBER THREAT LANDSCAPE

A. Definition and Scope

A cyber threat is defined as any malicious action that undermines the confidentiality, integrity, or availability of digital systems, networks, applications, or data. These threats may originate from various sources, including individual attackers, organized groups, nation-state actors, insiders, or automated malware. The motivations behind cyber threats are diverse, ranging from financial fraud and data theft to espionage, service disruption, and unauthorized access. The growing reliance on interconnected systems, cloud services, mobile devices, and Internet of Things (IoT) networks has enhanced communication and automation but has also expanded the potential for exploitation. The variety and persistence of these threats often exceed the capabilities of traditional security measures, underscoring the need for effective detection and mitigation strategies in modern security practices.

B. Principal Threat Categories

Malware remains a significant issue and includes various categories such as viruses, worms, Trojans, spyware, rootkits, and ransomware. Modern samples of malware employ strategies like obfuscation, polymorphism, and fileless execution to evade detection by signature-based solutions. Phishing continues to be widespread, primarily targeting human behaviour rather than technical vulnerabilities. The emergence of AI-generated content has complicated the task of identifying fraudulent communications, raising the challenge of distinguishing them from legitimate messages. Ransomware is particularly notable as one of the most economically damaging cybersecurity threats, with sophisticated campaigns that combine encryption, data exfiltration, and extortion, affecting industries such as healthcare, finance, government, education, and critical infrastructure. Additional significant threats include Distributed Denial of Service (DDoS) attacks, Advanced Persistent Threats (APTs), insider threats, botnets, credential theft, supply chain vulnerabilities, and zero-day exploits. Many of these threats are designed to bypass conventional security measures while maintaining persistent access, highlighting the need for detection systems that can recognize both established and emerging threat patterns[7], [20].

C. The Case for Intelligent Detection

The increasing scale and diversity of cybersecurity threats have accelerated the integration of artificial intelligence (AI) within the field. AI-driven approaches enable the automation of extensive security data analysis, facilitating the rapid detection of anomalous behaviour, classification of complex activities, and support for prompt responses. Consequently, AI plays a pivotal role in modern cyber defence, enhancing capabilities to identify, categorize, and react to an evolving threat landscape [24], [25].

IV. TRADITIONAL CYBER THREAT DETECTION

Prior to the widespread implementation of AI, cybersecurity relied on established detection methodologies that continue to serve as vital components of layered defence. These traditional techniques focus on recognizing known attack patterns, applying predefined rules, and monitoring for unusual behaviour. While effective against previously identified threats, they often encounter challenges when faced with sophisticated and rapidly evolving attacks.

A. Signature-Based Detection

Signature-based detection involves comparing files, network traffic, or system behaviour against a database of known attack signatures and identifying matches as malicious. This method boasts high accuracy and low false-positive rates for recognized threats; however, its effectiveness hinges on regular updates to the signature repository, leaving it vulnerable to zero-day exploits, polymorphic malware, and new variants that lack a corresponding signature [22], [26].

B. Rule-Based Detection

Rule-based detection employs expert-defined policies derived from known attack patterns, protocols, and organizational requirements, allowing for transparent decision-making and straightforward maintenance in stable environments. Nonetheless, it necessitates ongoing manual updates as attack strategies evolve, which limits its efficacy in large and dynamic networks. [26].

C. Anomaly-Based Detection

Anomaly-based detection focuses on identifying deviations from a learned baseline of normal behaviour, thereby flagging potentially unseen threats.

The primary challenge lies in defining what constitutes normal behaviour within complex environments where user activities and traffic fluctuate significantly, often resulting in elevated false-positive rates and increased workloads for analysts. [9], [21], [48].

D. Heuristic-Based Detection

Heuristic detection integrates predefined rules with behavioural analysis to evaluate suspicious characteristics, rather than relying solely on exact signatures. This approach enhances the detection of modified malware compared to traditional signature-based methods. Nevertheless, it may still miss highly advanced attacks and could produce false positives when legitimate actions resemble malicious behaviours.

E. Intrusion Detection and Prevention Systems

Conventional intrusion detection systems (IDS) and intrusion prevention systems (IPS) utilize these techniques to continuously monitor network traffic and system activity. Their efficacy is increasingly limited by factors such as encrypted communications, advanced persistent threats (APTs), AI-assisted attacks, and the escalating volume of security telemetry, which renders manual analysis and static rule applications impractical at scale [26].

F. Limitations and the Shift to Learning-Based Methods

These challenges have accelerated the integration of machine learning and deep learning technologies, which provide adaptive learning capabilities, automated feature extraction, enhanced scalability, and real-time detection of both known and emerging threats. Learning-based methods have become fundamental to the field of cybersecurity research and practice, though they come with their own set of complexities, as will be discussed in subsequent sections.

V. MACHINE LEARNING AND DEEP LEARNING FOR THREAT CLASSIFICATION

In contrast to static signatures and manual rules, learning-based methods analyze patterns within data and adjust to evolving attack behaviours. This adaptability enhances the capability to identify known threats, uncover new attack vectors, and enable quicker real-time responses. Consequently, AI has become fundamental to precise, scalable, and effective threat classification. Figure 2 presents a comprehensive taxonomy of the methods examined in this review.

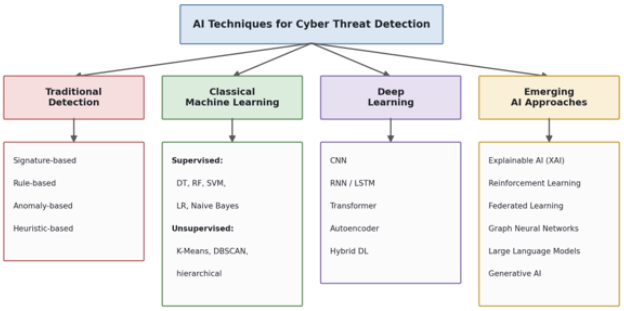


Fig. 2. Taxonomy of detection techniques covered in this review, from traditional methods to emerging AI approaches

A. Supervised Machine Learning

When labelled data is available, supervised methods such as decision trees, random forests, support vector machines (SVM), logistic regression, and naive Bayes are utilized to classify traffic and activities by learning the relationships between input features and attack categories. [13], [14], [15], [16]. These models demonstrate high accuracy for known attack types and are favourable for lightweight and interpretable applications. However, their performance is significantly influenced by the quality, diversity, and representativeness of the training data, and they may underperform when encountering patterns that are not represented in the training set [39].

B. Unsupervised Learning

In scenarios where labelled data is limited or non-existent, unsupervised methods can identify patterns and irregularities without relying on predefined categories. Clustering algorithms like K-Means, DBSCAN, and hierarchical clustering are designed to aggregate similar data points and highlight anomalies that could suggest potential attacks.

While these techniques can uncover new threats that were previously unrecognized, they tend to generate a higher rate of false positives compared to supervised methods, which constrains their effectiveness as standalone solutions in production environments. [21], [45].

C. Deep Learning

Deep learning enhances real-time classification by automatically extracting features from complex, high-dimensional security data. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and transformer models demonstrate robust performance in tasks such as intrusion detection, malware classification, phishing detection, and traffic analysis. [1], [10], [12]. Recurrent and convolutional architectures effectively capture temporal and spatial structures in traffic, while deep models have the capability to identify complex attack patterns that traditional methods often overlook. However, the deployment of these models is constrained by significant computational and data requirements, along with the need for meticulous optimization, particularly in environments with limited resources [40].

D. Hybrid Models

Hybrid approaches integrate machine learning and deep learning to mitigate the limitations inherent in individual models. Common configurations blend supervised learning with anomaly detection or deep learning with feature selection to enhance accuracy, robustness, and adaptability. These models have demonstrated potential in addressing complex, multi-stage attacks across various application domains; however, the added complexity may lead to increased tuning efforts and computational costs.

E. Determinants of Effectiveness

The efficacy of AI-based classification systems relies on factors such as data quality, feature engineering, interpretability, computational efficiency, and the ability to operate in real-time. As threats continue to evolve, models necessitate regular retraining and validation to maintain performance levels. Within these parameters, AI methods offer intelligent and adaptive detection capabilities that significantly outperform traditional static techniques. [6], [7], [10].

VI. EMERGING AI APPROACHES FOR CYBER THREAT DETECTION

A. Explainable Artificial Intelligence (XAI)

Many machine learning and deep learning models operate as opaque systems, hindering analysts' ability to comprehend the rationale behind classifying an input as malicious. In environments where prompt and justifiable decisions are critical, this lack of transparency can diminish trust and hinder investigations. Explainable AI (XAI) addresses this challenge by identifying the features, activities, or events that significantly influence predictions, enabling analysts to validate outcomes, distinguish real attacks from false positives, and full-fill accountability requirements in regulated fields such as healthcare, finance, and government [27], [28], [43], [44].

XAI has found applications in areas such as intrusion detection, malware identification, phishing detection, insider threat analysis, and traffic monitoring. Feature-attribution methods like SHAP assess the contribution of individual variables, while model-agnostic techniques provide explanations that are independent of the underlying algorithms. Nonetheless, two challenges persist: the balance between interpretability and predictive accuracy in highly complex models, and the lack of standardized metrics for evaluating the quality, stability, and usefulness of explanations. Despite these challenges, XAI plays a vital role in fostering transparent and reliable detection systems [44].

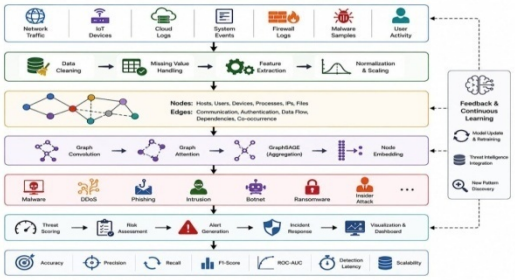


Fig. 3. Representative AI-driven cyber threat detection framework, illustrating data ingestion, graph-based representation, graph neural network processing, threat classification, response, and a continuous-learning feedback loop.

B. Reinforcement Learning

Reinforcement learning (RL) develops agents that enhance decision-making through interactions within an environment, relying on rewards and penalties instead of labelled data. This characteristic of RL makes it particularly effective for cybersecurity, where the nature of attack behaviour evolves over time. RL-based systems are capable of monitoring activities, evaluating threats, and determining defensive measures without the necessity for predefined signatures, thereby facilitating adaptive defences and minimizing the need for manual oversight. Use cases encompass intrusion detection, malware remediation, traffic management, cyber deception, vulnerability assessment, and automated incident response. Key challenges include the requirement for realistic training environments, prolonged training durations, meticulous reward structuring, maintaining stable learning processes, and ensuring resilience against adversarial manipulation, as well as achieving a balance between exploration and exploitation within extensive enterprise networks.

C. Federated Learning

Federated learning (FL) facilitates the training of a unified machine learning model across various devices or organizations while ensuring that raw data remains on-site. Participants share only model updates, making this approach particularly advantageous in cybersecurity contexts where data confidentiality is paramount. FL effectively mitigates data exposure risks and fosters secure collaboration among organizations. Its applications include intrusion detection, malware classification, anomaly detection, threat intelligence sharing, and privacy-centric analytics, particularly in Internet of Things (IoT) and edge computing environments [11], [29], [30]. However, FL faces several significant challenges, including the presence of non-independent and identically distributed (non-IID) data among participants, communication overhead, limitations posed by constrained edge devices, the need for model synchronization, and potential adversarial threats like model poisoning and inference attacks. Research is ongoing in areas such as secure aggregation and enhancing privacy safeguards [29], [30].

D. Graph Neural Networks (GNNs)

Graph neural networks (GNNs) process data organized in graph structures, where entities are represented as nodes and their relationships as edges. This framework is particularly relevant to cybersecurity, as aspects such as hosts, users, devices, files, and communications can naturally be modelled as graphs. GNNs leverage information from both the features of nodes and their connectivity to detect suspicious relationships, identify anomalies, and classify multi-stage attacks that conventional node-independent methods may overlook. Their applications span intrusion detection, malware analysis, fraud detection, threat intelligence, and attack prediction, proving to be effective against advanced persistent threats (APTs) that manifest through interconnected stages [31], [32]. However, GNNs face several limitations, including challenges in accurately constructing graphs from raw security data, significant computational requirements, and unresolved issues related to scalability, interpretability, and adversarial robustness within graph structures, as well as the capacity for real-time processing [32], [37].

E. Large Language Models (LLMs)

Large language models (LLMs) are designed to process extensive text data, demonstrating capabilities in natural language understanding, information extraction, and content generation. This functionality facilitates advancements in domains such as threat intelligence, security analysis, and decision support [34], [35]. Within the cybersecurity domain, LLMs are capable of analyzing unstructured data types, including threat reports, vulnerability descriptions, system logs, and incident histories. They can effectively extract indicators of compromise, summarize incidents, and provide support to security analysts. Research has indicated their utility in malware analysis, phishing detection, vulnerability assessment, report generation, and automated incident response. Given that numerous security-related parameters are conveyed through text and tabular formats, transformer-based models are well-suited for these applications. Studies show that transformer and LLM methodologies achieve competitive accuracy in intrusion detection tasks when evaluated against standard datasets [50]. Additionally, lightweight and privacy-preserving versions of these models have been proposed for use with constrained Internet of Things (IoT) and Industrial Internet of Things (IIoT) devices[51]. However, there are significant risks associated with LLMs, including high computational requirements, the potential for generating misleading outputs (hallucinations) that may impact decision-making, and vulnerability to adversarial inputs and prompt-injection attacks. Strategies such as retrieval-augmented generation, domain-specific fine-tuning, human oversight, and enhanced evaluation methods are being investigated to bolster the reliability of these systems[35], [37].

F. Generative AI

Generative AI generates new content, patterns, and synthetic data by leveraging learned distributions, which facilitates data augmentation, attack simulation, and automation in security processes. The use of synthetic samples can help mitigate dataset imbalance and privacy issues while enhancing the robustness and generalization of detection models. Key applications include automated vulnerability assessments, malware research, phishing detection, security code reviews, and support for incident response. However, these same capabilities also pose risks; adversaries may exploit generative models to create convincing phishing schemes, automate social engineering tactics, and develop malicious code. Concerns regarding data privacy, reliability, hallucination, and limited control over generated outputs necessitate responsible and secure deployment, alongside ongoing monitoring and governance [20].

Table 1 summarises the emerging approaches by working principle, advantages, challenges, and application.

TABLE I — Comparative summary of emerging AI approaches for cyber threat detection.

Approach	Working principle	Main advantage	Key challenge	Cybersecurity application
Explainable AI (XAI)	Produces interpretable explanations of model predictions	Transparency, analyst trust, accountability	No standard measure of explanation quality; accuracy trade-off	Threat analysis, monitoring, incident investigation
Reinforcement learning	Learns actions from environment feedback	Adaptive, automated response	Costly training; reward design; stability	Automated defence, dynamic response, security optimisation
Federated learning	Trains shared models without sharing raw data	Privacy preservation; collaboration	Non-IID data; communication cost; poisoning	Privacy-preserving IDS, distributed monitoring
Graph neural networks	Learns from relationships between entities	Detects relational and multi-stage attacks	Graph construction cost; scalability; robustness	Attack-path detection, network analysis, threat modelling
Large language models	Applies language understanding to security text	Contextual analysis; automation	High cost; hallucination; adversarial risk	Threat intelligence, vulnerability analysis, assistance
Generative AI	Generates content, scenarios, and synthetic data	Data augmentation; proactive testing	Validation and misuse risk	Attack simulation, reporting, security assistants

VII. BENCHMARK DATASETS

Public datasets play a crucial role in the advancement, assessment, and benchmarking of AI-driven detection models by supplying authentic network traffic, malicious activities, system behaviours, and attack scenarios within controlled environments. However, variations in data collection methods, attack classifications, feature representations, and traffic characteristics pose challenges for making direct comparisons of results across different studies [6], [10].

A. Established Intrusion Detection Datasets

Public datasets play a crucial role in the advancement, assessment, and benchmarking of AI-driven detection models by supplying authentic network traffic, malicious activities, system behaviours, and attack scenarios within controlled environments. However, variations in data collection methods, attack classifications, feature representations, and traffic characteristics pose challenges for making direct comparisons of results across different studies [3]. NSL-KDD enhanced the KDD Cup 1999 dataset by eliminating duplicates and balancing the distribution of attack types, establishing it as a widely used benchmark. However, it still lacks coverage of contemporary threats, including Advanced Persistent Threats (APTs), encrypted attacks, and modern malware[6].UNSW-NB15 offers an updated dataset that includes normal traffic alongside contemporary attack vectors, which encompass exploits, reconnaissance, and Denial of Service (DoS) attacks. It also provides an extensive array of flow features[4]. CICIDS2017 encompasses authentic benign and malicious traffic, reflecting a variety of attack types such as brute-force, botnet, DoS, web, and infiltration attacks. It includes flow features optimized for advanced modelling [8]. CSE-CIC-IDS2018 broadens this scope by incorporating a larger and more diverse set of enterprise traffic, facilitating scalable deep learning applications and real-time assessments[6].

B. IoT, IIoT, and Recent Datasets

The emergence of IoT and industrial IoT security has led to the development of specialized datasets. The BoT-IoT dataset offers IoT testbed traffic that includes both normal behaviour and various attack types such as DoS, DDoS, reconnaissance, and information theft, facilitating intrusion and anomaly detection [5].ToN-IoT combines network traffic, telemetry, operating system logs, and security events across both IoT and IIoT environments, enabling more comprehensive threat and anomaly analysis [41]. WUSTL-IIOT-2021 focuses on research pertinent to industrial IoT security [42], Additionally, the Edge-IIoTset enables evaluations of both centralized and federated learning approaches [11]. Recent datasets have emerged to address the limitations of earlier benchmarks. The CICIoT2023 dataset provides extensive real-time attack traffic gathered from a large variety of real IoT devices, encompassing attack types such as DDoS, reconnaissance, web, brute-force, spoofing, and those based on the Mirai botnet [52]. Table 2 presents a comparison of these datasets.

TABLE II — Benchmark and emerging datasets for AI-based cyber threat detection.

Dataset	Year	Environment	Representative attacks	Main limitation
KDD Cup 1999 [3]	1999	Simulated network	DoS, U2R, R2L, probe	Outdated patterns; redundant records
NSL-KDD [6]	2009	Simulated network	DoS, U2R, R2L, probe	No modern or encrypted attacks
UNSW-NB15 [4]	2015	Enterprise network	Exploits, reconnaissance, DoS, fuzzers	Benchmark may not reflect live traffic
CICIDS2017 [8]	2017	Enterprise network	Brute force, botnet, DoS, web, infiltration	May not cover all emerging threats
CSE-CIC-	2018	Enterprise	Brute force,	Large volume;

Dataset	Year	Environment	Representative attacks	Main limitation
IDS2018 [6]		network	DoS, botnet, infiltration	heavy preprocessing
BoT-IoT [5]	2019	IoT testbed	DoS, DDoS, reconnaissance, theft	IoT-specific; requires adaptation
ToN-IoT [41]	2020	IoT / IIoT	DoS, scanning, ransomware, injection	Heterogeneous sources complicate use
Edge-IIoTset [11]	2022	IoT / IIoT edge	Multiple IoT/IIoT attacks	Deployment and resource constraints
CICIoT2023 [52]	2023	Large-scale real IoT	DDoS, recon, web, brute force, spoofing, Mirai	High volume; still maturing in adoption

C. Dataset Limitations

Public datasets, while valuable, exhibit several inherent weaknesses, including class imbalance, inadequate representation of real incidents, outdated attack models, and restricted environmental diversity. Additionally, privacy constraints and challenges in gathering large-scale real traffic further limit their availability. Consequently, advancement in this field relies on the development of realistic, diverse, and consistently updated datasets that accurately represent operational conditions [6], [10].

VIII. EVALUATION METRICS

A thorough evaluation necessitates the use of multiple complementary metrics. It is essential to interpret reported values in conjunction with the associated dataset and protocol that generated them. The definitions provided below clarify that TP, TN, FP, and FN refer to true positives, true negatives, false positives, and false negatives, respectively.

A. Accuracy

Accuracy represents the ratio of correctly identified instances within a dataset. While it serves as a valuable overall metric, it can be misleading in the context of imbalanced datasets, where instances of normal traffic significantly exceed those of attacks. In such cases, a high accuracy value may obscure inadequate detection performance for less frequent classes.

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

B. Precision

Precision represents the proportion of flagged instances that are legitimate threats. High precision is crucial, as an abundance of false alerts can inundate analysts and hinder response times.

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

C. Recall

Recall, often referred to as sensitivity or the true positive rate, represents the proportion of actual threats that are successfully identified. Achieving a high level of recall is critical, as undetected attacks can lead to significant consequences.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

D. F1-Score

The F1-score represents the harmonic mean of precision and recall, offering significant insights into imbalanced datasets by effectively balancing the metrics of detection and false positives.

$$F1\ Score = 2 * \frac{Recall\ \text{X}\ Precision}{Recall + Precision}$$

E. False Positive Rate

The false positive rate represents the percentage of normal activities incorrectly identified as malicious. Reducing this rate is crucial for ensuring practical usability.

$$FPR = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

F. ROC-AUC, Latency, and Efficiency

The ROC curve establishes a relationship between the true positive rate and the false positive rate at various thresholds, with the area under the curve (AUC) serving as a summary of the model's discriminative power. Detection latency assesses the model's response time to incoming data, which is essential for real-time applications. Computational efficiency, encompassing aspects such as processing speed, memory usage, training duration, and scalability, plays a crucial role in determining the viability of deployment in large-scale or edge environments. Given that no single metric can encompass all relevant dimensions, a multi-metric evaluation approach is required for a comprehensive assessment of AI-based detection systems.

IX. COMPARATIVE SYNTHESIS

This section synthesizes the reviewed studies in a collective manner rather than in sequential order. It identifies areas of consensus and divergence within the literature, explores the reasons for these differences, and highlights outstanding questions. Table 3 presents dataset-focused studies with verifiable contributions, while Table 4 outlines foundational methods that recapitulate throughout the discipline. Table 5 compiles the thematic synthesis.

A. Consensus: AI Outperforms Traditional Detection

Surveys and empirical studies demonstrate that AI techniques, particularly deep learning, surpass signature-based and rule-based detection methods when addressing complex and novel threats. This is primarily due to their capability to automatically extract features and adapt to evolving patterns[10], [24], [25]. This observation is consistent across various domains, including intrusion detection, malware classification, and IoT security, and is applicable regardless of the specific architecture employed. However, it is important to note that this consensus is nuanced: the advantages of AI methods are most evident with complex, high-dimensional data and less prominent in cases of simple, well-defined attacks, where traditional methods continue to perform competitively.

B. Tension: Accuracy Against Cost and Deployability

There exists an inherent tension between predictive performance and practical cost in machine learning approaches. Advanced techniques such as deep learning, graph neural networks, and transformer models enhance the ability to detect complex patterns; however, they necessitate increased computational resources, larger datasets, and more extensive parameter tuning [1], [10], [50]. In contrast, classical machine learning techniques offer compelling advantages for lightweight and interpretable applications. Comparative analyses of IoT and IIoT datasets indicate that well-selected lightweight models can perform competitively against more complex models when operating under resourceconstraints [19], [39]. Consequently, the debate surrounding the efficacy of various methodologies is less about identifying an absolute superior approach and more about determining the optimal operating point. The choice of method should be guided by specific data characteristics, latency requirements, and available resources, rather than by a universal ranking system.

C. The Comparability Problem

Reported accuracies present challenges for comparison due to variations in datasets, feature sets, class-imbalance management, and evaluation methodologies [6], [10]. Many commonly utilized benchmarks have reached saturation, leading to near-ceiling accuracies that may indicate the characteristics of the benchmark rather than the true operational challenges, a concern highlighted in critical analyses of machine learning applications for network intrusion detection[49].

Inconsistencies in handling class imbalance exacerbate this issue, as accuracy metrics derived from skewed data can mask inadequate detection of infrequent classes [36]. These aspects illustrate why aggregate figures across different studies should be interpreted cautiously and emphasize the need for accompanying dataset characteristics with any reported results. Thus, Table 3 focuses on verifiable contributions rather than performance metrics that may not be comparable across different contexts.

TABLE III — Verified dataset-centered studies relevant to AI-based cyber threat detection.

Study	Contribution	Dataset	Key consideration
Tavallaee et al. [3]	Analysis of benchmark limitations	KDD Cup 99	Legacy benchmark; simulated traffic
Moustafa and Slay [4]	Modern IDS dataset	UNSW-NB15	May not fully reflect live networks
Koroniotis et al. [5]	Realistic IoT testbed dataset	BoT-IoT	IoT settings need continuous adaptation
Sharafaldin et al. [8]	Benign and multi-attack traffic	CICIDS2017	May not capture all emerging threats
Ferrag et al. [10]	Deep-learning IDS survey	Multiple	Dataset heterogeneity limits comparison
Ferrag et al. [11]	Centralised and federated evaluation	Edge-IIoTset	Deployment and resource constraints
Ismail et al. [19]	Lightweight ML comparison	ToN_IoT, WUSTL-IIOT, Edge-IIoTset	Results remain dataset-dependent
Neto et al. [52]	Large-scale real-time IoT dataset	CICIoT2023	High volume; adoption still maturing

D. Promise and Uncertainty in Emerging Approaches

Emerging methodologies in the field exhibit promising potential, yet they are variably validated. Explainable AI (XAI) enhances trust and facilitates investigation; however, it is hampered by the absence of universally accepted metrics for assessing explanation quality and can compromise accuracy for greater interpretability [27], [28], [43], [44]. Federated learning offers a solution to privacy concerns in decentralized environments but encounters challenges such as non-IID data, communication overhead, and vulnerability to poisoning attacks [29], [30]. Graph neural networks effectively model relational structures and can address multi-stage threats like Advanced Persistent Threats (APTs), albeit with trade-offs in graph construction, computational demands, and robustness issues [31], [32], [33]. Transformer and large language model (LLM)-based techniques proficiently manage unstructured threat intelligence as well as text and tabular security data.

Recent literature indicates that these approaches achieve competitive detection accuracy; nonetheless, challenges such as hallucination, susceptibility to adversarial attacks, and high computational costs impede their reliability [34], [35], [50], [51A recurring theme across these methodologies is the juxtaposition of significant reported potential against a backdrop of limited independent and operational validation.

TABLE IV — Foundational AI and machine learning methods relevant to cyber threat detection.

Source	Contribution	Relevance to detection	Key consideration
Goodfellow et al. [1]	Deep learning foundations	Representation learning for complex data	High computation and data demand
Russell and Norvig [2]	General AI foundations	Concepts for intelligent decision-making	Not cybersecurity-specific
Chen and Guestrin [13]	XGBoost	Efficient boosting for classification	Depends on features and tuning
Breiman [14]	Random forest	Strong tabular baseline for IDS	Needs retraining as data shifts
Cortes and Vapnik [15]	Support vector machine	Established supervised classifier	Scaling to large data is hard
Pedregosa et al. [16]	Scikit-learn	Standard ML implementations	A library, not a detection study
Goodfellow et al. [17]	Adversarial examples	Robustness risk for AI defence	Needs dedicated evaluation
Humayun et al. [20]	AI and IoT security survey	Links AI adoption with IoT risk	Broad survey scope

E. Recurring Methodological Limitations

The literature consistently identifies four significant limitations. First, the heavy dependence on a limited number of benchmark datasets restricts external validity [6], [10]. Second, the majority of evaluations are conducted offline, with few studies documenting live deployment or long-term operational performance [11], [49]. Third, the use of inconsistent metrics and inadequate handling of class imbalance diminishes the comparability of results [36]. Fourth, the assessment of robustness against adversarial manipulation is frequently conducted weakly or omitted entirely [17], [37], [38]. These limitations are cumulative: collectively, they indicate that strong performance on benchmarks does not assure operational reliability.

TABLE V — Thematic synthesis of the reviewed literature.

Theme	Consensus	Tension or disagreement	Open issue
AI vs traditional detection	AI improves detection of complex and unseen threats	Advantage smallest for simple, known attacks	When simpler methods suffice
Model selection	Deep and graph models raise accuracy	Cost and deployability favour lightweight models	Best operating point per setting
Evaluation	Multi-metric assessment is needed	Reported accuracies are not comparable	Standard, reproducible protocols
Emerging approaches	XAI, FL, GNN, and LLM are promising	Independent validation is limited	Operational, live evidence
Robustness and privacy	Both are recognised as critical	Often tested weakly or omitted	Systematic robustness and privacy testing

F. Research Gap

The collected evidence indicates a notable deficiency. Existing research has yet to yield standardized, reproducible, and live-validated assessments of explainable and resource-efficient AI utilizing contemporary, representative datasets, and the ability to generalize across diverse environments has not been established. Addressing this deficiency necessitates the development of methods that are accurate, explainable, scalable, resource-efficient, and adaptable, with evaluation conducted in conditions that closely mimic operational networks rather than relying on saturated benchmarks. This deficiency directly informs the challenges and directions discussed in the subsequent sections.

X. CHALLENGES AND LIMITATIONS

A. Data Quality and Class Imbalance

Detection quality is inherently reliant on the quality of the training and evaluation datasets used. Numerous public datasets are characterized by outdated attack patterns, duplicates, incomplete features, and imbalanced class distributions, which constrains the models' ability to generalize effectively to real-world conditions. In scenarios where malicious activity constitutes only a small percentage of the data, models often exhibit a bias towards typical traffic, resulting in the oversight of rare yet significant attacks. Therefore, there is a continuous necessity for diverse, balanced, and frequently updated datasets[6], [10], [36].

B. Adversarial Attacks and Security Risks

AI models present inherent vulnerabilities as potential attack surfaces. Adversarial perturbations can lead to misclassification, while threats such as data poisoning, model evasion, and unauthorized access can compromise their reliability[17], [37]. A standardized taxonomy of attacks and mitigations for machine learning has been established, highlighting the necessity for explicit evaluation of robustness [38]. Enhancing robustness is crucial for ensuring reliable detection.

C. Interpretability and Computational Cost

A considerable number of high-accuracy deep models exhibit a lack of transparency, which complicates the validation process during security assessments. Additionally, these models require substantial computational resources, memory, and power, leading to increased deployment costs and limiting their applicability in resource-constrained environments. Achieving an equilibrium between accuracy, interpretability, and efficiency continues to be a significant challenge[1], [10], [12].

D. *Privacy, Scalability, and Real-Time Deployment*

AI-powered systems handle sensitive traffic and user data, which introduces privacy and regulatory considerations during the collection and analysis phases. These systems need to function effectively within extensive and rapidly evolving networks, maintaining accuracy while reducing latency, optimizing resources, and integrating seamlessly with existing infrastructure. It is crucial to address privacy, scalability, and real-time requirements for successful implementation[20].

E. *Summary*

Artificial Intelligence has enhanced detection capabilities via intelligent classification, automated analysis, and real-time decision-making. However, challenges such as data quality, adversarial robustness, interpretability, efficiency, privacy, scalability, and deployment remain significant barriers to practical implementation. Addressing these challenges is essential for the development of reliable and trustworthy systems.

XI. RESEARCH GAPS

A. *Scarcity of Realistic Datasets*

Benchmark datasets have significantly contributed to the field; however, many fail to accurately reflect contemporary networks or the rapid evolution of attack strategies, as they are predominantly generated in controlled environments. The need for realistic, continuously updated, and openly accessible datasets continues to be a priority [6], [10].

B. *Limited Generalisation*

Models are frequently trained and evaluated using a singular dataset or environment, which constrains their ability to transfer. Variations in architecture, user behaviour, and attack patterns influence performance; therefore, enhancing generalization across a variety of settings is crucial [6], [10], [19].

C. *Insufficient Explainability and Trust*

Achieving high accuracy alone is insufficient for practical application in operational settings. Numerous models provide limited interpretability, complicating the validation of results for analysts. There is a need for enhanced transparency and standardized explanation methodologies to facilitate informed decision-making[27], [28], [44].

D. *Weak Real-World Validation*

Most research primarily assesses performance using offline benchmarks, with limited testing conducted in operational environments. Factors such as network scale, changing behaviors, resource constraints, and ongoing data streams influence actual performance; therefore, live deployment and long-term evaluation are essential[11], [49].

E. *Summary*

The literature indicates notable advancements; however, it reveals deficiencies in realistic data, generalization, explainability, and operational validation. Focusing on these areas will enhance the reliability, scalability, and trustworthiness of cyber-defense systems, as illustrated in Figure 4 and Table 6.

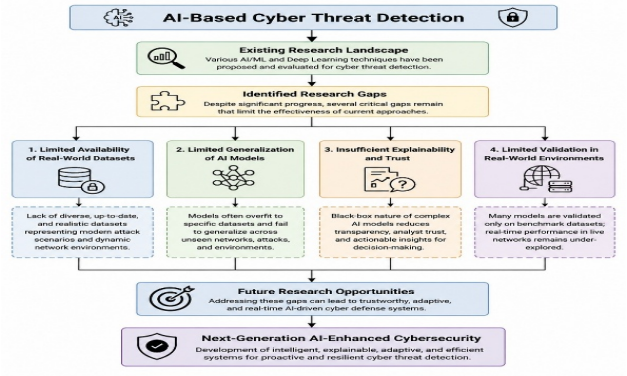


Fig. 4. Research gaps and future research opportunities in AI-based cyber threat detection.

TABLE VI — Key research gaps and recommended priorities.

Research domain	Current limitation	Effect on detection	Recommended direction
Benchmark datasets	Poor representation of modern attacks and dynamic networks	Reduced reliability on real networks	Build realistic, continually updated, diverse datasets
Model adaptability	Trained for narrow settings; weak on unseen patterns	Lower performance in heterogeneous networks	Design adaptive models for evolving threats
Explainability	Limited visibility into predictions	Reduced analyst confidence	Integrate XAI to support decisions
Operational validation	Reliance on offline benchmarks	Real-world effectiveness uncertain	Evaluate in live environments over time
Resource efficiency	High compute requirements	Costly, limited edge use	Develop lightweight, efficient models

XII. FUTURE DIRECTIONS

A. *Explainable and Trustworthy AI*

Future systems must integrate accuracy with clear explanatory frameworks. Transparency enhances analyst confidence, facilitates regulatory compliance, and assists in investigation and response efforts [27], [28], [44].

B. *Federated and Privacy-Preserving Learning*

Privacy requirements support the adoption of federated learning, a method that enables the training of shared models without the need to exchange sensitive data. Efforts should focus on enhancing communication efficiency, robustness, and secure aggregation to facilitate privacy-conscious detection in distributed environments[11], [29], [30].

C. *Edge AI and IoT Security*

The expansion of Internet of Things (IoT) and edge computing introduces constraints in resource availability and increases the vulnerability landscape. Implementing lightweight, efficient models for detection at the edge can minimize latency, enhance scalability, and bolster security in distributed environments[11], [20], [51].

D. *Generative AI and Large Language Models*

Generative AI and large language models (LLMs) facilitate applications in various domains, including threat intelligence, automated reporting, malware analysis, vulnerability assessment, and awareness. Ongoing research is necessary to tackle potential misuse by adversaries while ensuring that defensive systems remain secure, reliable, and resilient to manipulation[34], [35], [50].

E. *Autonomous and Adaptive Defence*

Future systems are anticipated to transition from passive detection methods to adaptive defense mechanisms that evolve by learning from emerging attack vectors, updating their knowledge base, and executing responses with minimal human intervention. The integration of adaptive learning, threat intelligence, and automated response strategies will facilitate the development of resilient architectures suited for complex environments[11], [20].

F. Summary

Progress will rely on solutions that are explainable, preserve privacy, adapt to varying conditions, and operate efficiently. Developments in federated learning, edge AI, large language models, and autonomous defence systems are anticipated to enhance real-time detection capabilities and reinforce next-generation infrastructures. Figure 5 and Table 7 summarize these directions.

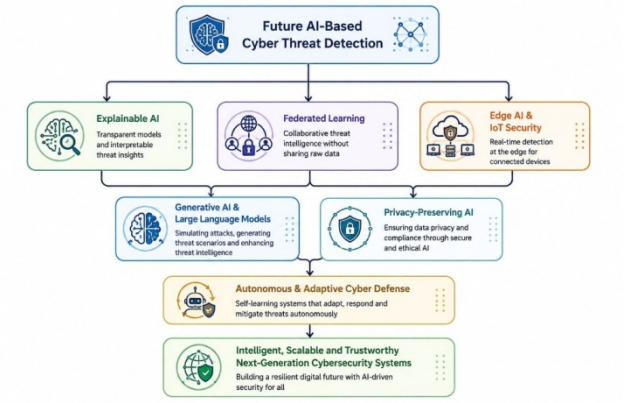


Fig. 5. Roadmap for future AI-based cyber threat detection

TABLE VII — Future research directions in AI-based cyber threat detection.

Future area	Current challenge	Expected outcome
Explainable AI	Limited transparency of decisions	Better interpretability, trust, and accountability
Federated learning	Privacy risk in collaborative training	Secure decentralised learning without data sharing
Edge AI and IoT	Resource limits and real-time constraints	Lightweight, efficient models for edge devices
Generative AI and LLMs	Dual-use risk and AI-assisted attacks	Intelligent analysis with automated security operations
Autonomous defence	Speed and complexity of attacks	Adaptive, self-learning systems with automated response

XIII. CONCLUSION

Artificial intelligence has emerged as a pivotal component in cybersecurity, facilitating intelligent, adaptive, and real-time threat detection. This review assesses AI-driven threat classification methodologies, encompassing traditional detection techniques, classical machine learning, deep learning, and newer approaches. It includes an analysis of benchmark datasets, evaluation metrics, comparative insights, challenges, gaps, and future directions.

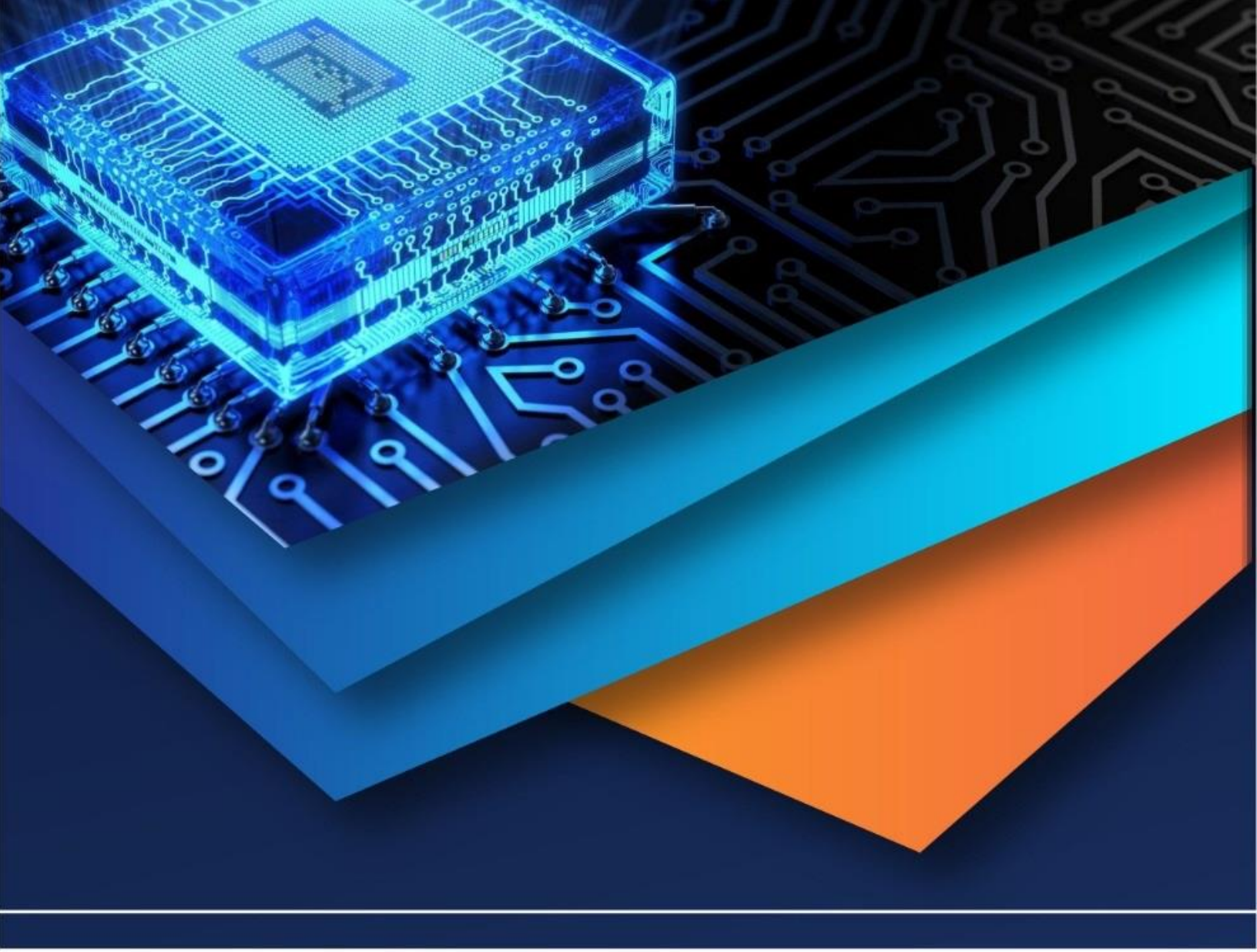
The findings indicate that AI methodologies enhance the accuracy, speed, and scalability of detection in comparison to traditional methods, with machine learning, deep learning, explainable AI, federated learning, and edge intelligence contributing to more effective and adaptive defence mechanisms. However, significant limitations remain, including a lack of realistic datasets, challenges in interpretability, susceptibility to adversarial attacks, high computational demands, privacy concerns, and inadequate validation under real-world conditions. The comparative analysis reveals that while strong benchmark results are observed, they have not yet translated into reliable operational performance.

Future research should thus focus on the development of explainable, secure, resource-efficient, and adaptive models, which must be assessed using contemporary datasets and in practical environments. Emphasis on privacy-preserving learning, autonomous defence mechanisms, and standardized, reproducible evaluation processes will likely influence the evolution of next-generation cyber-defence systems. Results from individual studies should be interpreted within the context of their specific datasets, settings, and evaluation metrics.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[2] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Hoboken, NJ, USA: Pearson, 2020.
[3] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD Cup 99 data set," in Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl. (CISDA), 2009, pp. 1-6, doi: 10.1109/CISDA.2009.5356528.
[4] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in Proc. Mil. Commun. Inf. Syst. Conf. (MilCIS), 2015, pp. 1-6, doi: 10.1109/MilCIS.2015.7348942.
[5] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," Future Gener. Comput. Syst., vol. 100, pp. 779-796, 2019, doi: 10.1016/j.future.2019.05.041.
[6] M. Ring, D. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," Comput. Secur., vol. 86, pp. 147-167, 2019, doi: 10.1016/j.cose.2019.06.005.
[7] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," Cybersecurity, vol. 2, no. 20, 2019, doi: 10.1186/s42400-019-0038-7.
[8] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. 4th Int. Conf. Inf. Syst. Secur. Privacy (ICISSP), 2018, pp. 108-116, doi: 10.5220/0006639801080116.
[9] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," IEEE Commun. Surveys Tuts., vol. 16, no. 1, pp. 303-336, 2014, doi: 10.1109/SURV.2013.052213.00046.
[10] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," J. Inf. Secur. Appl., vol. 50, art. 102419, 2020, doi: 10.1016/j.jisa.2019.102419.
[11] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning," IEEE Access, vol. 10, pp. 40281-40306, 2022, doi: 10.1109/ACCESS.2022.3165809.
[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015, doi: 10.1038/nature14539.
[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2016, pp. 785-794, doi: 10.1145/2939672.2939785.
[14] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.
[15] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273-297, 1995, doi: 10.1007/BF00994018.
[16] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.
[17] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learn. Representations (ICLR), 2015.
[18] D. Dua and C. Graff, "UCI machine learning repository," Univ. California, Irvine, CA, USA, 2019.
[19] S. Ismail, S. Dandan, and A. Qushou, "Intrusion detection in IoT and IIoT: Comparing lightweight machine learning techniques using TON_IoT, WUSTL-IIOT-2021, and Edge-IIoTset datasets," IEEE Access, vol. 13, pp. 73468-73485, 2025.
[20] M. Humayun, N. Tariq, M. Alfayad, M. Zakwan, G. Alwakid, and M. Assiri, "Securing the Internet of Things in artificial intelligence era: A comprehensive survey," IEEE Access, vol. 12, pp. 25469-25490, 2024, doi: 10.1109/ACCESS.2024.3365634.
[21] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surv., vol. 41, no. 3, art. 15, 2009, doi: 10.1145/1541880.1541882.
[22] D. E. Denning, "An intrusion-detection model," IEEE Trans. Softw. Eng., vol. SE-13, no. 2, pp. 222-232, 1987, doi: 10.1109/TSE.1987.232894.
[23] M. V. Mahoney and P. K. Chan, "An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection," in Proc. Recent Advances in Intrusion Detection (RAID), 2003, pp. 220-237, doi: 10.1007/978-3-540-45248-5_13.
[24] A. H. Salem, S. M. Azzam, O. E. Emam, and A. A. Abohany, "Advancing cybersecurity: A comprehensive review of AI-driven detection techniques," J. Big Data, vol. 11, art. 105, 2024, doi: 10.1186/s40537-024-00957-y.
[25] A. Ali, M. Charfeddine, B. Ammar, B. B. Hamed, F. Albalwy, A. Alqarafi, and A. Hussain, "Unveiling machine learning strategies and considerations in intrusion detection systems: A comprehensive survey," Front. Comput. Sci., vol. 6, art. 1387354, 2024, doi: 10.3389/fcomp.2024.1387354.
[26] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," J. Netw. Comput. Appl., vol. 36, no. 1, pp. 16-24, 2013, doi: 10.1016/j.jnca.2012.09.004.
[27] C. Molnar, Interpretable Machine Learning, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
[28] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.
[29] J. L. Hernandez-Ramos, G. Karopoulos, E. Chatzoglou, V. Kouliaridis, E. Marmol, A. Gonzalez-Vidal, and G. Kambourakis, "Intrusion detection based on federated learning: A systematic review," arXiv:2308.09522, 2023.

[30] P. Kairouz et al., "Advances and open problems in federated learning," Found. Trends Mach. Learn., vol. 14, nos. 1-2, pp. 1-210, 2021, doi: 10.1561/2200000083.
[31] F. Guan, T. Zhu, W. Zhou, and K.-K. R. Choo, "Graph neural networks: A survey on the links between privacy and security," Artif. Intell. Rev., vol. 57, art. 40, 2024, doi: 10.1007/s10462-023-10656-4.
[32] M. Zhong, M. Lin, C. Zhang, and Z. Xu, "A survey on graph neural networks for intrusion detection systems: Methods, trends and challenges," Comput. Secur., vol. 141, art. 103821, 2024, doi: 10.1016/j.cose.2024.103821.
[33] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proc. Int. Conf. Learn. Representations (ICLR), 2017.
[34] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, T. Yu, Y. Liu, and H. Wang, "Large language models for cyber security: A systematic literature review," arXiv:2405.04760, 2024.
[35] J. Zhang, H. Bu, H. Wen, Y. Liu, H. Fei, R. Xi, L. Li, Y. Yang, H. Zhu, and D. Meng, "When LLMs meet cybersecurity: A systematic literature review," Cybersecurity, vol. 8, no. 1, art. 55, pp. 1-41, 2025, doi: 10.1186/s42400-025-00361-w.
[36] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263-1284, 2009, doi: 10.1109/TKDE.2008.239.
[37] G. Li, P. Zhu, J. Li, Z. Yang, N. Cao, and Z. Chen, "Security matters: A survey on adversarial machine learning," arXiv:1810.07339, 2018.
[38] A. Vassilev, A. Oprea, A. Fordyce, H. Anderson, X. Davies, and M. Hamin, "Adversarial machine learning: A taxonomy and terminology of attacks and mitigations," NIST AI 100-2e2023, Nat. Inst. Standards Technol., 2024, doi: 10.6028/NIST.AI.100-2e2023.
[39] A. B. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Commun. Surveys Tuts., vol. 18, no. 2, pp. 1153-1176, 2016, doi: 10.1109/COMST.2015.2494502.
[40] A. Aldhaheri, F. Alwahedi, M. A. Ferrag, and A. Battah, "Deep learning for cyber threat detection in IoT networks: A review," Internet Things Cyber-Phys. Syst., vol. 4, pp. 110-128, 2024, doi: 10.1016/j.iotcps.2023.09.003.
[41] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," IEEE Access, vol. 8, pp. 165130-165150, 2020, doi: 10.1109/ACCESS.2020.3022862.
[42] M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan, and R. Jain, "WUSTL-IIOT-2021 dataset for IIoT cybersecurity research," IEEE DataPort, 2022, doi: 10.21227/yftq-n229.
[43] A. Rjoub, J. Bentahar, O. Abdel Wahab, R. Mizouni, A. Song, R. Cohen, H. Otrok, and A. Mourad, "A survey on explainable artificial intelligence for cybersecurity," arXiv:2303.12942, 2023.
[44] C. Mendes and T. N. Rios, "Explainable artificial intelligence and cybersecurity: A systematic literature review," arXiv:2303.01259, 2023.
[45] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," arXiv:1901.03407, 2019.
[46] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," IEEE Trans. Emerg. Topics Comput. Intell., vol. 2, no. 1, pp. 41-50, 2018, doi: 10.1109/TETCI.2017.2772792.
[47] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," IEEE Access, vol. 5, pp. 21954-21961, 2017, doi: 10.1109/ACCESS.2017.2762418.
[48] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," Comput. Netw., vol. 51, no. 12, pp. 3448-3470, 2007, doi: 10.1016/j.comnet.2007.02.001.
[49] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. IEEE Symp. Secur. Privacy, 2010, pp. 305-316, doi: 10.1109/SP.2010.25.
[50] H. Kheddar, "Transformers and large language models for efficient intrusion detection systems: A comprehensive survey," Inf. Fusion, vol. 124, art. 103347, 2025, doi: 10.1016/j.inffus.2025.103347.
[51] M. A. Ferrag, M. Ndhlovu, N. Tihanyi, L. C. Cordeiro, M. Debbah, T. Lestable, and N. S. Thandi, "Revolutionizing cyber threat detection with large language models: A privacy-preserving BERT-based lightweight model for IoT/IIoT devices," IEEE Access, vol. 12, pp. 23733-23750, 2024, doi: 10.1109/ACCESS.2024.3363469.
[52] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment," Sensors, vol. 23, no. 13, art. 5941, 2023, doi: 10.3390/s23135941.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY