Ai-Driven Threat Detection: A Structured Review of Algorithmic, Cloud-Native, and Human-Centered Security Approaches

Sujata Gawade
VPM's College of Engineering, Thane

***Abstract*—The rapid growth of automated, polymorphic, and multi-vector cyber threats has exposed the limitations of traditional signature-based network defense. Artificial intelligence, machine learning, and deep learning have therefore become central to proactive, data-driven threat detection. This paper presents a structured review of AI-driven threat detection across three dimensions: algorithmic adaptability, deployment architecture, and threat surface evolution.**

Three representative studies are examined: an algorithmic taxonomy of support vector machine, random forest, convolutional neural network, and generative adversarial network models; a cloud-native detection architecture evaluated on a simulated enterprise network; and a user-centered framework for AI-generated synthetic threats such as phishing and deepfakes.

The review compares these studies in terms of architecture, evaluation environment, metrics, operational advantages, and deployment bottlenecks. The analysis identifies four major barriers to production adoption: black-box opacity, latency and resource constraints at the edge, adversarial vulnerabilities including poisoning and evasion, and false-positive-driven alert fatigue.

Finally, the paper outlines future research directions, including neuro-symbolic AI, adversarial-aware model compression, automated adversarial retraining loops, and human-in-the-loop explainable interfaces. The central finding is that production-ready AI-driven threat detection requires not only accurate models, but also transparent, scalable, adversarial robust, and human-aware defensive architectures.

***Index Terms*—AI-driven threat detection, machine learning, deep learning, explainable AI, adversarial machine learning, cloud-native security, cybersecurity, human-in-the-loop security.**

I. INTRODUCTION

The rapid growth and increasing sophistication of cyber threats have transformed network defense into an adversarial race against automation. Traditional signature-based and rule-based architectures remain foundational to cybersecurity, but they struggle with zero-day vulnerabilities, polymorphic malware, adaptive intrusion campaigns, and multi-vector breaches. These conventional systems also suffer from structural latency: security analysts usually extract indicators of compromise and write static rules only after an attack has already been observed or has successfully bypassed existing controls.

Artificial intelligence, machine learning, and deep learning have emerged as essential paradigms for intelligent, data-centric threat mitigation. By analyzing high-velocity telemetry from endpoints, network traffic, cloud infrastructure, and user behavior, AI-driven systems can identify latent behavioral anomalies that may indicate malicious activity before escalation. Classical machine learning models such as support vector machines and random forests can operate effectively on structured telemetry, while deep learning models such as convolutional neural networks and recurrent architectures can analyze unstructured, sequential, or high-dimensional security data. However, as defensive AI evolves, so does the threat landscape. Attackers increasingly use automation, adversarial manipulation, and generative technologies to improve the scale, realism, and evasiveness of cyberattacks. Reports of malicious or abuse-oriented generative tools, including systems marketed as WormGPT, FraudGPT, and PoisonGPT, illustrate how generative AI can lower the barrier for phishing, malware adaptation, synthetic identity fraud,

and social engineering. Although the availability and sophistication of such tools vary, they highlight the need for defensive systems that account for AI-assisted adversaries.

AI-driven threat detection therefore creates a dual challenge. On one hand, defenders need adaptive models that can detect unknown or evolving attacks. On the other hand, these models must be interpretable, computationally feasible, resilient to adversarial manipulation, and usable by human security analysts. A detection model with high accuracy on benchmark datasets may still fail in production if it produces excessive false positives, introduces unacceptable latency, cannot explain its decisions, or becomes vulnerable to poisoning and evasion.

Existing research often examines AI-driven cyber defense through isolated lenses: model accuracy, infrastructure scalability, or user susceptibility to synthetic attacks. Production-ready threat detection, however, requires these dimensions to be studied together. This paper therefore reviews AI-driven threat detection across three interrelated axes: algorithmic adaptability, deployment architecture, and threat surface vectors. This review contributes to the literature in three ways. First, it integrates algorithmic, architectural, and human-centered perspectives that are often treated separately in AI-driven cybersecurity research. Second, it compares representative studies across evaluation environment, metrics, operational benefits, and deployment bottlenecks. Third, it identifies cross-cutting barriers to production adoption and proposes near-term research directions for robust, explainable, and resource-efficient AI-enabled threat detection.

II. LITERATURE REVIEW

The reviewed literature is organized along three dimensions: algorithmic adaptability, deployment architecture, and threat surface vectors. These dimensions are interdependent. Algorithm choice constrains what can be deployed in real time; deployment architecture determines which threats can be inspected at scale; and the threat surface increasingly includes human users as well as technical systems.

A. Algorithmic Adaptability and Technical Diversity

Dhanushkodi and Thejas [1] survey AI-enabled threat detection and show that algorithm choice is closely tied to the structure of input data. Their taxonomy identifies broad use of support vector machines, random forests, convolutional neural networks, generative adversarial networks, and explainable AI overlays. Three major algorithmic families are especially relevant.

First, classical machine learning models such as support vector machines and random forests remain valuable because of their deterministic behavior, relatively low computational cost, and strong performance on structured data [2], [3]. They are commonly used for tabular telemetry such as netflow records, packet headers, access logs, authentication events, and system-level indicators. Their main advantage is operational simplicity, but their effectiveness depends heavily on feature engineering and may degrade when adversaries modify attack behavior.

Second, deep learning models such as convolutional neural networks, recurrent neural networks, and long short-term memory networks are useful for high-dimensional, unstructured, or sequential data [4], [5]. For example, CNNs can process malware binaries transformed into two-dimensional image-like representations, allowing spatial pattern recognition to detect obfuscated malware variants. Sequential models can analyze temporal attack chains, command sequences, or network flows. However, deep models typically require more data, more compute, and more careful tuning than classical machine learning approaches.

Third, generative and decentralized paradigms have become increasingly important. Generative adversarial networks can produce synthetic threat-like samples, potentially improving training diversity when real zero-day examples are scarce. However, synthetic data raises concerns about realism, label reliability, and overfitting to artificial patterns. Federated learning enables distributed model training across edge devices or organizational nodes without requiring raw logs to be centralized. This improves privacy but introduces communication overhead, synchronization problems, and vulnerability to poisoned local updates.

Overall, the algorithmic literature shows that there is no universally optimal model for threat detection. Model selection depends on data type, latency requirements, explainability needs, adversarial risk, and deployment environment.

B. Deployment Architecture and Runtime Infrastructure

Moving from algorithmic design to production deployment requires scalable execution frameworks. Farzaan et al. [6] propose a cloud-native implementation that wraps machine learning engines in containerized infrastructure. Their architecture uses AWS EC2 node clusters orchestrated by Kubernetes so that detection services can scale horizontally with the velocity of incoming traffic. This deployment model uses micro-segmentation to separate high-throughput traffic capture from compute-intensive inference engines. This design helps prevent latency spikes during heavy traffic analysis or Distributed Denial of Service validation sweeps. The system also uses hybrid machine learning engines with dynamic contamination ratios for real-time traffic filtering, deployed as containerized microservices.

Cloud-native deployment offers important benefits. It improves scalability, modularity, maintainability, and fault isolation. Detection components can be updated independently, replicated under load, and monitored through cloud-native observability tools. However, this approach also introduces new operational complexity. Kubernetes clusters require careful security configuration, container images must be hardened, service-to-service communication must be protected [7], and orchestration misconfigurations can themselves become attack vectors [8].

Runtime performance is also critical. In production security environments, detection quality cannot be evaluated by accuracy alone. CPU utilization, memory consumption, packet-capture drop rate, alert latency, and false-positive rate directly affect whether a system is usable by security teams.

C. Threat Surface Vectors: Autonomous Defense and Synthetic Attackers

Traditional cybersecurity systems classify technical anomalies such as privilege escalation, buffer overflows, lateral movement, suspicious authentication patterns, command-and-control traffic, and malicious payload execution. However, Almarwani et al. [9] show that the modern threat perimeter increasingly includes human-targeted, AI-synthesized deception. Generative AI enables hyper-personalized phishing, synthetic voice impersonation, deepfake video attacks [10], fake identity creation, and automated social engineering at scale. These attacks do not always rely on malware or anomalous network behavior. Instead, they exploit trust, urgency, authority, and cognitive overload [11]. Therefore, threat surfaces can no longer be assessed through a purely technical lens. A unified AI-driven cybersecurity framework must combine automated technical security with user-centered protective behavior. Automated systems can classify suspicious messages, detect synthetic media indicators, and flag abnormal account behavior, but human users still play a central role in recognizing deception, following verification procedures, and resisting manipulation. Human-in-the-loop security is therefore not optional; it is a necessary component of resilient cyber defense.

III. METHODOLOGY

This paper uses a structured narrative review methodology to examine representative AI-driven threat detection approaches. The purpose is not to conduct a large-scale statistical meta-analysis, but to synthesize the relationships among algorithms, deployment environments, and emerging human-centered threat surfaces.

A. Search Strategy

The selection process targeted peer-reviewed literature indexed in IEEE Xplore and IEEE Access because these sources provide scientific authority, technical relevance, and verifiable publication metadata. The search covered three overlapping conceptual domains: core AI methods, threat landscape, and specialized deployment environments.

The following Boolean expressions were used against titles, abstracts, and index keywords:

- “Artificial Intelligence” OR “Machine Learning” OR “Deep Learning” AND “Threat Detection” OR “Cyber Threat Mitigation” OR “Zero-Day Anomaly”
- “AI-Powered” OR “Machine Learning” AND “Cloud Infrastructure” OR “WIDS” OR “Containerized Operations” OR “Kubernetes”
- “Generative AI” OR “Synthetic Threats” OR “Adversarial Attacks” AND “Cybersecurity Framework” OR “Protective Behavior”

The review focused on recent literature addressing AI-enabled cybersecurity, including algorithmic models, cloud-native defensive systems, and human-centered responses to generative cyber threats.

B. Inclusion and Exclusion Criteria

Studies were included if they satisfied the following criteria:

- The study was published in a peer-reviewed journal, conference, or official academic repository.
- The study focused primarily on AI-driven threat detection, cyber threat mitigation, synthetic cyber threats, or cybersecurity decision support.
- The study provided algorithmic, architectural, empirical, or human-centered analysis relevant to AI-based security.

The study included reproducible or clearly described elements, such as model families, neural configurations, evaluation environments, datasets, testbeds, or user-study procedures.

Studies were excluded if they met any of the following conditions:

- The work was a dissertation, non-peer-reviewed white paper, commercial opinion piece, or informal technical commentary.
- The study lacked mathematical, empirical, architectural, or algorithmic grounding.
- The work addressed general IT operations without explicit threat-classification or cybersecurity relevance.
- The study discussed AI in general terms without a clear connection to cyber defense, adversarial risk, or threat detection.

C. Primary Data Targets

Three baseline papers were selected to span the operational range of AI-driven cybersecurity:

- An algorithmic and theoretical taxonomy of AI-enabled threat detection [1].
- A cloud-native infrastructure and implementation study validated on a simulated enterprise network of 250 endpoints, 25 servers, and 15 network devices [6].
- A human-centered study of AI-driven synthetic threats, user awareness, detection, and protective behavior [9].

These three works were selected because they represent distinct but complementary layers of AI-driven cybersecurity: model-level detection, infrastructure-level deployment, and human-targeted adversarial risk. This selection enables the review to examine not only detection accuracy, but also runtime feasibility, explainability, adversarial robustness, and user-level vulnerability.

The study-selection framework is summarized in Fig. 1. The figure organizes the review corpus into three analytical profiles: an algorithmic profile, a deployment profile, and an adversarial or human-centered profile.

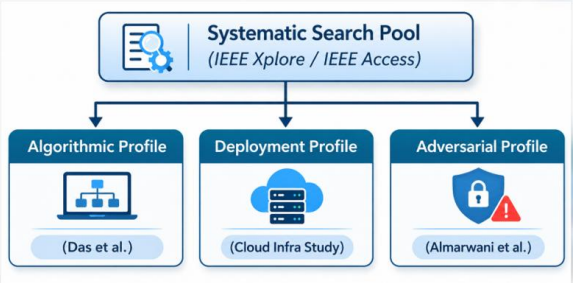


Fig. 1. Study selection framework showing how the review corpus was organized into three analytical profiles: algorithmic adaptability, deployment architecture, and adversarial or human-centered threat surfaces.

IV. RESULTS AND DISCUSSION

A. Thematic Taxonomy

The reviewed literature separates into three major dimensions: algorithmic adaptability, deployment architecture, and threat surface vectors. These categories interact rather than stand alone. Algorithm choice constrains the type of telemetry that can be analyzed; deployment architecture determines whether analysis can occur at production speed; and threat surfaces increasingly include human deception, not only malware and network intrusion.

Fig. 2 presents the proposed taxonomy of AI-driven threat detection. The first layer concerns the detection model itself, including classical machine learning, deep learning, generative AI, and explainable AI. The second layer concerns where and how the model is deployed, including edge environments, cloud-native systems, Kubernetes clusters, containerized services, and micro-segmented infrastructure. The third layer concerns the monitored threat surface, including malware, zero-day exploits, synthetic deception, phishing, and deepfakes.

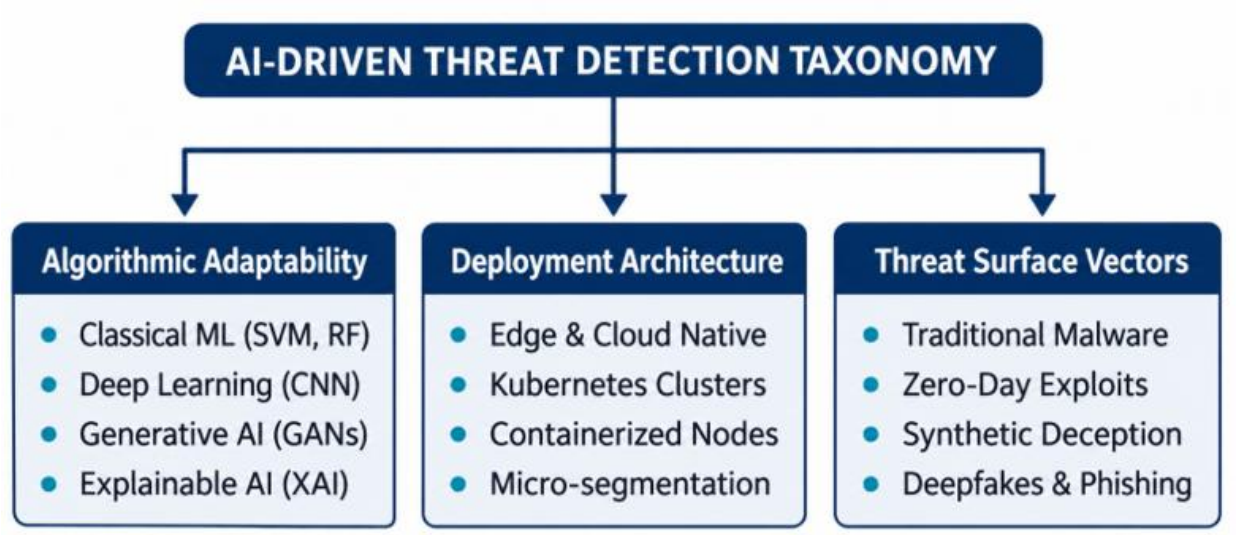


Fig. 2. Taxonomy of AI-driven threat detection organized around algorithmic adaptability, deployment architecture, and threat surface vectors.

B. Comparative Analysis of the Reviewed Studies

Table I compares the three reviewed studies across their core focus, AI methods, evaluation environment, metrics, advantages, bottlenecks, explainability, adversarial robustness, and production readiness.

Table I. Comparison of the reviewed studies

Parameter	Algorithmic Taxonomy [1]	Cloud Deployment Study [6]	User-Centered Framework [9]
Core focus	Profiling, classification metrics, and categorization of AI, ML, and DL models in cyber defense.	Systems engineering, runtime scalability, and false-positive reduction in enterprise cloud environments.	Human-centric analysis of user responses to AI-generated synthetic deception.
Methods / AI engine	Taxonomy of SVM, RF, CNN, GAN, and XAI-based approaches.	Hybrid ML engines with dynamic contamination ratios for real-time traffic filtering, deployed as containerized microservices.	Behavioral validation combining verification technologies, automated classification, and human cognitive modeling.
Evaluation environment	Meta-analytical evaluation over public datasets such as NSL-KDD and CICIDS2017.	Simulation of 250 endpoints, 25 servers, and 15 network devices on AWS EC2 nodes.	Empirical user study of responses to AI-generated phishing and deepfakes.
Key metrics	TPR, AUC, latency variation, explainability scores.	FPR, CPU usage, memory use under load, packet-capture drop rate, alert latency.	Recognition rate, competence gap, time-to-detection, protective-behavior adherence.
Main advantage	Provides a blueprint for matching data types to suitable models and supports transparent model selection.	Reduces alert fatigue by lowering false positives without reducing true-positive accuracy.	Bridges automated protection and user vulnerabilities with a defense design for the human element.
Main bottleneck	Vulnerable to poisoning and evasion; limited validation metrics for training data quality.	High architectural complexity and compute overhead; possible packet drops during burst traffic.	Depends on continuous user training; lacks automated enforcement to fully override human error.
Explainability support	Discusses XAI overlays, but interpretability depends on model type.	Limited operational explainability unless paired with analyst-facing dashboards.	Strong human-centered orientation, but less focused on technical model transparency.

Adversarial robustness	Recognizes poisoning and evasion risks but does not fully solve them.	Improves operational resilience but remains exposed to poisoned telemetry and API probing.	Addresses synthetic deception but relies heavily on user competence and behavior.
Production readiness	Useful for model selection and conceptual design.	High relevance to enterprise cloud environments.	Useful for awareness, training, and policy design; less mature as an automated enforcement system.
Reproducibility	Depends on dataset availability and model configuration details.	Moderate; testbed size and cloud architecture are described, but implementation details affect replication.	Depends on participant sample, scenario design, and measurement instruments.

C. Metrics Used in AI-Driven Threat Detection

Detection performance is commonly measured using true positive rate, false positive rate, precision, recall, F1-score, and area under the receiver operating characteristic curve. These metrics evaluate whether a system can correctly identify malicious and benign activity.

Operational performance requires additional measures, including alert latency, CPU utilization, memory consumption, packet-capture drop rate, and system throughput. These metrics are essential because an accurate model may still be impractical if it consumes excessive resources or causes dropped packets during high traffic.

Human-centered studies require different metrics. These include recognition rate, time-to-detection, protective-behavior adherence, user confidence, and competence gap. Such metrics are especially relevant for AI-generated phishing, deepfakes, and social engineering attacks, where technical indicators may be insufficient.

D. The Black-Box Dilemma and Interpretability

High-performing deep learning models such as CNNs, RNNs, and LSTMs often return anomaly labels or risk scores without a human-readable audit trail. This creates operational difficulty because analysts may not know whether an alert corresponds to a low-and-slow zero-day exploit, a benign configuration change, or a temporary deviation from historical baselines.

Black-box opacity slows incident response. Analysts must inspect model inputs, correlate logs, and reconstruct attack context before taking action. This delay is especially problematic in high-speed environments where containment decisions must be made quickly.

Opacity also creates compliance and governance risk. Organizations increasingly need to explain automated security decisions, especially when those decisions affect user access, account suspension, transaction blocking, or incident escalation. Post-hoc explanation methods such as SHAP [12] and LIME [13] can improve interpretability, but they add computational overhead and provide approximations rather than complete accounts of model logic.

A key direction for future research is therefore explainable AI designed specifically for cybersecurity operations. Analyst-facing explanations should translate anomaly vectors into operational language, such as unusual authentication sequence, abnormal data-transfer volume, rare process-tree behavior, or deviation from peer-group access norms.

E. Resource Constraints at the Edge

Deep learning models with millions of parameters often require substantial memory, high-performance CPUs, or GPU acceleration. Edge nodes such as IoT gateways, industrial controllers, mobile devices, and remote sensors usually have limited power, memory, and compute capacity. Running a heavy detection model on such devices can create latency spikes, packet drops, and incomplete telemetry capture.

This creates a latency-security trade-off. A more complex model may improve detection accuracy, but if it delays inspection or drops traffic, it can weaken the overall security posture. Lightweight models, quantization, pruning, and split inference can help reduce computational burden.

Federated learning offers another approach by allowing distributed devices to train collaboratively without sharing raw telemetry. This improves privacy and reduces the need for centralized data collection. However, federated learning introduces communication overhead, synchronization challenges, and vulnerability to poisoned updates from compromised clients.

F. Adversarial Vulnerabilities

Machine learning models depend on statistical patterns in training data, which means they expose an attack surface of their own [14]. Adversaries can attack the model, the training pipeline, the telemetry source, or the inference interface.

The first major risk is evasion. In evasion attacks, adversaries apply small, calculated modifications to malware binaries, packet features, or behavioral sequences. These perturbations preserve malicious functionality while shifting the model's classification toward benign [15]. This is especially dangerous when the model relies on brittle statistical features rather than robust semantic understanding.

The second risk is data poisoning. During continuous retraining, an insider or compromised sensor can inject malicious telemetry labeled as benign. Over time, this shifts decision boundaries and creates persistent blind spots. Poisoning is particularly concerning in cloud-native and federated settings where large volumes of telemetry are collected from many sources.

The third risk is model inversion, probing, and reconnaissance. If adversaries can repeatedly query exposed security APIs, they may infer decision thresholds, identify feature sensitivities, or approximate the model's behavior. This enables the design of future attacks that are more likely to bypass detection.

Addressing these threats requires adversarial testing, trusted data provenance, robust training, anomaly detection for training data, secure model update pipelines, and rate-limited access to inference APIs.

G. False-Positive Burden and Alert Fatigue

AI models detect statistical variance, but they do not always understand operational context. A legitimate database backup by a new engineer may resemble data exfiltration. A software update may resemble suspicious process execution. A sudden traffic spike may indicate either a DDoS attack or a legitimate business event.

High false-positive rates create alert fatigue. Security teams may receive thousands of alerts per day, many of which are low-value or redundant. Over time, analysts become desensitized and may dismiss genuine low-signal threats as noise. This is especially dangerous for advanced persistent threats, which often unfold slowly and subtly.

The cloud-native approach in [6] directly targets this issue by reducing false positives without reducing true-positive accuracy. However, false-positive reduction should not be treated only as a model-tuning task. It also requires contextual enrichment, asset criticality scoring, user behavior baselines, threat intelligence correlation, and analyst feedback loops.

H. Synthesis of Findings

The reviewed studies show that AI-driven threat detection cannot be evaluated by accuracy alone. A model with strong benchmark performance may fail in production if it introduces latency, lacks interpretability, produces excessive false positives, or collapses under adversarial perturbation. Similarly, a scalable cloud-native detector may still be insufficient against synthetic social engineering if it does not incorporate human-centered risk signals.

The central challenge is therefore not simply model optimization. The real challenge is designing trustworthy detection ecosystems that integrate algorithmic robustness, operational scalability, explainable decision-making, and human interpretability. Classical machine learning remains useful for lightweight structured telemetry. Deep learning improves detection of complex and high-dimensional patterns. Cloud-native systems improve scalability. Human-centered frameworks address social and cognitive attack surfaces. The strongest future systems will combine these capabilities rather than treating them as separate domains.

V. RESEARCH GAPS AND FUTURE DIRECTIONS

Table II summarizes the major barriers identified in the review and maps them to future research directions.

Table II Research gaps and future directions

Challenge	Current Limitation	Future Research Direction
Black-box opacity	Deep models provide limited auditability and weak analyst trust.	Neuro-symbolic AI and contextual XAI dashboards.

Edge constraints	Heavy models cause latency, packet drops, and high-power consumption.	Quantization, pruning, lightweight architectures, and split inference.
Adversarial evasion	Perturbed malware or traffic can bypass classifiers.	Robust training, adversarial testing, and semantic feature modeling.
Data poisoning	Continuous retraining can absorb malicious or mislabeled telemetry.	Data provenance, trusted pipelines, poisoning detection, and secure update validation.
Model probing	Exposed APIs can reveal model thresholds and decision behavior.	Query monitoring, rate limiting, model hardening, and privacy-preserving inference.
Alert fatigue	High false-positive rates overwhelm analysts.	Context-aware prioritization, risk scoring, and analyst feedback loops.
Synthetic deception	Users struggle to detect AI-generated phishing and deepfakes.	Human-in-the-loop interfaces, behavioral analytics, and verification workflows.

A. Neuro-Symbolic AI

Neuro-symbolic AI combines neural pattern recognition with symbolic reasoning. In cybersecurity, this can allow systems to learn from large volumes of telemetry while also enforcing deterministic security rules. For example, a symbolic layer could block actions that violate known security invariants even if a neural model is uncertain or fooled by adversarial perturbation.

Neuro-symbolic systems can also improve explainability. Instead of producing only a risk score, the system can explain that an alert was triggered because a user accessed an unusual resource, from an unusual location, during an abnormal time window, using a device that recently showed suspicious process behavior.

B. Resource-Optimized Edge Models

Future AI-driven threat detection must operate across cloud, endpoint, IoT, and industrial environments. This requires models that are both accurate and lightweight. Adversarial-aware quantization can reduce precision from FP32 to INT8 or INT4 while preserving robustness. Dynamic neural pruning can disable inactive pathways according to traffic volume or device constraints.

Research should focus not only on compressing models, but also on evaluating whether compressed models remain robust under evasion, poisoning, and distribution shift.

C. Automated Adversarial Retraining Loops

Cloud-native digital twins of enterprise networks can be used to simulate attacks before they occur in production. In such environments, offensive AI agents can continuously probe, evade, and poison defensive models. The resulting attack traces can be used to retrain models, patch weaknesses, and update detection logic before real attackers exploit the same vulnerabilities.

However, automated retraining loops must be carefully governed. Without trusted validation, they may amplify poisoned data or overfit to simulated attacks. Secure model governance, dataset versioning, and human approval thresholds are therefore required.

D. Human-in-the-Loop Explainable Interfaces

Human analysts remain central to cybersecurity. Future systems should provide natural-language explanations, visual attack timelines, confidence scores, recommended actions, and links to supporting evidence. Effective dashboards should reduce cognitive load rather than simply display more alerts.

Human-in-the-loop explainable interfaces are especially important for synthetic social engineering, where users and analysts must evaluate semantic cues, behavioral inconsistencies, and contextual risk. AI systems should support human judgment through verification workflows, not replace it blindly.

VI. LIMITATIONS OF THIS REVIEW

This review is limited by its reliance on three representative studies rather than a large-scale quantitative meta-analysis. While the selected studies span algorithmic, deployment, and human-centered dimensions, they do not fully capture all AI-driven threat detection domains, such as industrial control systems, vehicular networks, blockchain security, endpoint detection and response platforms, or operational technology environments.

In addition, differences in datasets, metrics, and experimental environments prevent direct statistical comparison of model performance. Public benchmark

datasets may not fully reflect modern enterprise networks, encrypted traffic, cloud-native workloads, or adversarial adaptive threat actors [16]–[18]. The cloud deployment study provides practical architectural insight, but simulation environments may not capture all production variability. The human-centered framework highlights user vulnerability to synthetic threats, but its findings depend on participant sample, scenario realism, and measurement design.

Future reviews should expand the corpus, include multiple academic databases, apply standardized quality-assessment criteria, and compare results across more diverse datasets and deployment settings.

VII. CONCLUSION

AI-driven cyber defense has shifted threat detection from static signature matching toward adaptive, behavior-based anomaly mitigation. The reviewed literature shows that effective deployment depends on more than model accuracy. Algorithms must be interpretable, computationally feasible, resilient to adversarial manipulation, and integrated with human-centered defense mechanisms.

Classical machine learning models remain useful for structured telemetry and resource-constrained environments. Deep learning models improve detection of complex, high-dimensional, and evolving threats. Generative and federated approaches expand the training and deployment possibilities, but they also introduce risks related to synthetic data quality, poisoned updates, and adversarial manipulation. Cloud-native architectures improve scalability and operational flexibility, but they add infrastructure complexity and require careful management of latency, false positives, and packet capture reliability.

The threat surface is also expanding beyond technical indicators. AI-generated phishing, deepfakes, synthetic identity attacks, and personalized deception require security systems that account for human behavior and cognitive vulnerability. A purely technical detection model is insufficient when the attack vector targets trust and decision-making.

Future AI-enabled security systems should therefore combine statistical learning with symbolic reasoning, adversarially robust retraining, resource-aware model optimization, and explainable human-in-the-loop interfaces. Such hybrid architectures are essential for building defensive systems that are not only accurate, but also transparent, scalable, resilient, and trustworthy under continuous adversarial pressure.

REFERENCES

[1]. K. Dhanushkodi and S. Thejas, “AI Enabled Threat Detection: Leveraging Artificial Intelligence for Advanced Security and Cyber Threat Mitigation,” IEEE Access, vol. 12, pp. 173127–173136, Nov. 2024, doi: 10.1109/ACCESS.2024.3493957.

[2]. A. L. Buczak and E. Guven, “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection,” IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, Second Quarter 2016, doi: 10.1109/COMST.2015.2494502.

[3]. A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, “Survey of Intrusion Detection Systems: Techniques, Datasets and Challenges,” Cybersecurity, vol. 2, no. 1, Art. no. 20, 2019, doi: 10.1186/s42400-019-0038-7.

[4]. M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, “Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study,” Journal of Information Security and Applications, vol. 50, Art. no. 102419, Feb. 2020, doi: 10.1016/j.jisa.2019.102419.

[5]. R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, “Deep Learning Approach for Intelligent Intrusion Detection System,” IEEE Access, vol. 7, pp. 41525–41550, 2019, doi: 10.1109/ACCESS.2019.2895334.

[6]. M. A. M. Farzaan, M. Ghanem, A. El Hajjar, and D. N. Ratnayake, “AI-Powered System for an Efficient and Effective Cyber Incidents Detection and Response in Cloud Environments,” IEEE Transactions on Machine Learning in Communications and Networking, vol. 3, pp. 623–643, Apr. 2025, doi: 10.1109/TMLCN.2025.3564912.

[7]. National Institute of Standards and Technology, “Zero Trust Architecture,” NIST Special Publication 800-207, Gaithersburg, MD, USA, 2020, doi: 10.6028/NIST.SP.800-207.

[8]. S. Sultan, I. Ahmad, and T. Dimitriou, “Container Security: Issues, Challenges, and the

Road Ahead," IEEE Access, vol. 7, pp. 52976–52996, 2019, doi: 10.1109/ACCESS.2019.2911732.

[9]. R. Almarwani, M. Almarwani, and F. Almarwani, "AI-Driven Synthetic Threats in Cybersecurity: A User-Centered Framework for Awareness, Detection, and Protective Behavior," IEEE Access, vol. 14, pp. 41737–41752, Mar. 2026, doi: 10.1109/ACCESS.2026.3673980.

[10]. Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," ACM Computing Surveys, vol. 54, no. 1, pp. 1–41, Jan. 2021, doi: 10.1145/3425780.

[11]. R. Dhamija, J. D. Tygar, and M. Hearst, "Why Phishing Works," in Proc. SIGCHI Conference on Human Factors in Computing Systems, Montreal, Canada, 2006, pp. 581–590, doi: 10.1145/1124772.1124861.

[12]. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. Advances in Neural Information Processing Systems, Long Beach, CA, USA, 2017, pp. 4765–4774.

[13]. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust you? Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[14]. N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "SoK: Security and Privacy in Machine Learning," in Proc. IEEE European Symposium on Security and Privacy, Paris, France, 2018, pp. 399–414, doi: 10.1109/EuroSP.2018.00035.

[15]. B. Biggio and F. Roli, "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning," Pattern Recognition, vol. 84, pp. 317–331, Dec. 2018, doi: 10.1016/j.patcog.2018.07.023.

[16]. N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," in Proc. Military Communications and Information Systems Conference, Canberra, Australia, 2015, pp. 1–6, doi: 10.1109/MilCIS.2015.7348942.

[17]. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in Proc. 4th International Conference on Information Systems Security and Privacy, Funchal, Portugal, 2018, pp. 108–116, doi: 10.5220/0006639801080116.

[18]. M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A Survey of Network-Based Intrusion Detection Data Sets," Computers & Security, vol. 86, pp. 147–167, Sep. 2019, doi: 10.1016/j.cose.2019.06.005.