Journal of Science Research International (JSRI)
ISSN: 2456-6365 (Online)

OPEN ACCESS

MACHINE LEARNING FOR CYBER THREAT DETECTION: LIMITATIONS, DEPLOYMENT BARRIERS, AND EMERGING RESEARCH PATHWAYS

Sudha Ramesh, Ashwin Ambalathara*,

Sonal Bansod, Anirudh Menon and Kumari Sanjana

Department of Computer Science,

Pillai College of Arts, Commerce and Science (Empowered Autonomous), New Panvel

**Corresponding author E-mail: aashwin23cs@student.mes.ac.in*

Received: 13 January 2026	Revised: 13 February 2026	Accepted: 29 March 2026	Published: 17 April 2026

DOI: https://doi.org/10.5281/zenodo.19625075

Abstract:

Machine learning has become an important component of modern cybersecurity systems due to its ability to analyse large volumes of security data and detect complex attack patterns. As cyber threats continue to evolve, many organizations have adopted learning-based intrusion detection and threat monitoring solutions. Despite strong performance reported in academic evaluations, machine learning–based security systems often fail to achieve comparable results in real-world environments. This paper presents a systematic review of recent peer-reviewed literature to examine the key limitations affecting their practical deployment. The review identifies persistent challenges related to adversarial robustness, model interpretability, data quality, zero-day attack detection, cross-domain generalization, computational scalability, and privacy preservation. These challenges are analyzed as interconnected barriers rather than isolated technical issues. The findings highlight a disconnect between laboratory-based evaluation methods and operational cybersecurity requirements, and they outline research directions focused on robustness, transparency, and deployment-oriented design.

Keywords: *Machine Learning, Cybersecurity, Intrusion Detection, Adversarial Robustness, Explainable AI, Privacy-Preserving Security,*

1. Introduction

The rapid expansion of digital technologies has fundamentally transformed modern society by enabling unprecedented levels of connectivity, automation, and data exchange. Organizations increasingly rely on cloud computing, distributed systems, mobile platforms, and Internet of Things (IoT) devices to support critical

operations. While these technologies improve efficiency and scalability, they also expand the attack surface and increase exposure to cyber threats. Adversaries exploit system complexity, software vulnerabilities, and human factors to compromise networks, exfiltrate sensitive data, and disrupt essential services. Traditional cybersecurity mechanisms, including signature-based intrusion detection systems and rule-based firewalls, were designed for comparatively static threat environments. These approaches depend on predefined attack signatures and manually crafted rules, which limit their effectiveness against novel or evolving threats. As cyber attackers increasingly employ automation and obfuscation techniques, static defense mechanisms struggle to provide timely and reliable protection.

Machine learning has emerged as a promising alternative, enabling security systems to learn patterns of normal and malicious behavior directly from data. Learning-based models can adapt to changing environments, detect subtle anomalies, and scale to large and complex datasets. Consequently, machine learning techniques have been widely adopted for tasks such as intrusion detection, malware classification, and behavioral analysis. Numerous studies report high detection accuracy when evaluating these models on benchmark datasets.

However, strong performance in experimental settings does not always translate to effective real-world deployment. Machine learning-based security systems often experience significant performance degradation when exposed to adversarial manipulation, evolving attack strategies, and operational constraints. Challenges related to data quality, model interpretability, cross-environment generalization, and computational scalability frequently limit their practical effectiveness.

This paper argues that the primary challenges facing machine learning–based cyber threat detection are not solely algorithmic but systemic in nature. By reviewing recent peer-reviewed literature, this study examines the interconnected limitations that hinder the reliable deployment of these systems in operational environments. Understanding these limitations is essential for guiding future research toward machine learning solutions that are not only accurate but also robust, interpretable, and deployable within real-world cybersecurity infrastructures.

2. Literature review

The application of machine learning to cybersecurity has attracted significant research interest due to the increasing scale, complexity, and sophistication of modern cyber threats. Machine learning techniques have been explored across a wide range of security tasks, including network intrusion detection, malware analysis, phishing detection, botnet identification, and user behavior modeling. Early studies primarily relied on traditional supervised learning algorithms such as support vector machines, decision trees, naïve Bayes classifiers, and ensemble methods. These approaches demonstrated measurable improvements over rule-based systems when evaluated under controlled conditions.

In recent years, deep learning techniques have gained prominence in cybersecurity research. Convolutional neural networks, recurrent neural networks, autoencoders, and deep reinforcement learning models have been applied to model high-dimensional network traffic and complex behavioral patterns. Many studies report high detection accuracy and reduced reliance on manual feature engineering. However, these performance gains are typically demonstrated using benchmark datasets or simulated environments, raising concerns regarding their external validity and real-world applicability.

A recurring issue in the literature is the heavy dependence on outdated, synthetic, or narrowly scoped datasets. Widely used datasets such as KDD Cup 99 and NSL-KDD remain prevalent despite well-documented shortcomings,

including redundant records, unrealistic traffic distributions, and limited representation of modern attack techniques. Although newer datasets such as UNSW-NB15, CICIDS, and IoT-specific datasets have been introduced, studies indicate that they still fail to capture the full diversity and evolution of real-world cyber threats. As a result, models trained on these datasets often exhibit inflated accuracy and poor generalization when deployed in operational environments.

Another significant body of research focuses on the vulnerability of machine learning models to adversarial manipulation. Adversarial machine learning studies demonstrate that attackers can evade detection by subtly modifying input features without altering malicious functionality. In cybersecurity contexts, such evasion strategies exploit the learned decision boundaries of intrusion detection systems. While this problem has been extensively explored in computer vision, its treatment in cybersecurity is more complex due to protocol-level constraints and functional requirements. Existing defense mechanisms often assume static adversaries and are rarely evaluated against adaptive or resource-aware attackers, limiting their practical effectiveness.

Model interpretability and transparency have also emerged as critical concerns in the literature. Many high-performing machine learning models operate as black boxes, providing limited explanations for their predictions. Several studies emphasize that this lack of transparency undermines analyst trust and complicates incident response in security operations centers. Explainable artificial intelligence techniques, including feature attribution methods and post hoc explanation models, have been proposed to improve interpretability. However, their adoption in real-time cybersecurity systems remains limited, and few studies empirically evaluate their impact on analyst decision-making or operational efficiency.

Data availability and quality issues further constrain the development and deployment of machine learning–based cybersecurity solutions. Security datasets are often highly imbalanced, with benign traffic significantly outnumbering malicious samples. In addition, labeled attack data are scarce, costly to obtain, and frequently incomplete or noisy. Organizational concerns related to confidentiality and privacy further restrict data sharing, resulting in fragmented datasets that limit model robustness. Techniques such as federated learning, differential privacy, and secure multi-party computation have been proposed as potential solutions, but empirical evaluations of their scalability and effectiveness in cybersecurity contexts remain limited.

The literature also highlights challenges related to generalization and adaptability. Numerous studies report that models trained in one organizational or network environment perform poorly when deployed in different contexts. Variations in network topology, traffic patterns, system configurations, and threat landscapes contribute to this performance degradation. These challenges are particularly pronounced in Internet of Things and edge computing environments, where resource constraints and heterogeneity further complicate deployment.

Finally, several studies identify operational and scalability challenges that are often overlooked in academic research. Real-time detection requirements, high data throughput, and computational limitations can significantly affect model performance in production systems. In addition, high false-positive rates remain a persistent concern, as excessive alerts can overwhelm security analysts and reduce the effectiveness of automated detection systems. These operational factors underscore the gap between experimental success and real-world deployment.

Overall, the existing literature demonstrates the potential of machine learning to enhance cyber threat detection but reveals a fragmented research landscape that prioritizes detection accuracy over deployment feasibility. Relatively few studies adopt holistic evaluation frameworks that consider adversarial robustness, interpretability, data realism, generalization, scalability, and human factors simultaneously. This review addresses this gap by

synthesizing these limitations as interconnected challenges that must be addressed to enable reliable, trustworthy, and deployable machine learning–based cybersecurity systems.

3. Methodology

This study employs a structured literature review approach to examine recent research on machine learning–based cyber threat detection. Peer-reviewed journal articles and conference proceedings published between 2019 and 2026 were collected from established academic databases, including IEEE Xplore, the ACM Digital Library, and Scopus. These sources were selected to ensure coverage of high-quality and relevant research in cybersecurity and machine learning.

The literature search utilized multiple combinations of keywords such as machine learning, cybersecurity, intrusion detection, adversarial attacks, explainable AI, zero-day detection, federated learning, and privacy-preserving security. Studies were included if they applied machine learning techniques to cybersecurity problems and explicitly addressed system limitations, real-world deployment challenges, or future research directions.

Selected studies were analyzed using qualitative thematic analysis. Findings related to robustness, interpretability, data quality, scalability, generalization, and privacy preservation were extracted and organized into recurring themes. This approach enabled the identification of consistent patterns across the literature, with emphasis on system-level challenges rather than isolated technical improvements.

4. Results

The analysis identified seven major categories of limitations impacting machine learning–based cyber threat detection systems.

4.1 Adversarial vulnerability

Adversarial vulnerability was identified as a major limitation in a substantial portion of the reviewed studies. Multiple works demonstrated that machine learning–based intrusion detection systems are susceptible to evasion attacks, in which adversaries deliberately manipulate input features to bypass detection while preserving malicious behavior. These attacks exploit the learned decision boundaries of models and can significantly reduce detection rates. Although some studies proposed defensive mechanisms, such as adversarial training or feature randomization, empirical evaluations under realistic and adaptive threat models were limited. As a result, robustness against adversarial behavior remains insufficiently validated.

4.2 Limited model interpretability

Limited interpretability emerged as a recurring concern, particularly for deep learning–based approaches. Many studies reported that high-performing models provide minimal insight into the factors influencing their predictions. This opacity complicates alert validation and incident response processes in operational environments. Several papers noted that security analysts are less likely to trust or act upon alerts generated by systems that cannot provide meaningful explanations. While explainable artificial intelligence techniques were occasionally incorporated, their evaluation was often restricted to qualitative demonstrations rather than systematic assessment of operational impact.

4.3 Data quality and availability challenges

Data-related limitations were among the most frequently cited issues in the literature. The reviewed studies consistently reported challenges associated with class imbalance, noisy labels, and the limited availability of labeled attack data. Many models were trained on benchmark datasets that do not reflect contemporary network traffic or evolving attack techniques. Additionally, some studies highlighted that dataset-specific artifacts can lead

to overfitting, resulting in inflated performance metrics that do not generalize beyond the evaluation environment. These data deficiencies were found to negatively affect both detection accuracy and model robustness.

4.4 Zero-Day and novel attack detection

The detection of zero-day and previously unseen attacks remains an unresolved challenge. Anomaly-based and unsupervised learning approaches were commonly proposed as solutions; however, the reviewed literature indicated that these methods often generate high false-positive rates when deployed in realistic environments. Several studies reported that distinguishing between benign behavioral changes and genuinely malicious activity remains difficult. Consequently, zero-day detection systems frequently require extensive tuning and human oversight to remain operationally viable.

4.5 cross-environment generalization

Poor generalization across different deployment environments was widely reported. Models trained on data from a specific network or organizational context often experienced significant performance degradation when evaluated on data from other environments. Differences in network architecture, traffic composition, user behavior, and threat profiles were identified as contributing factors. Few studies employed cross-dataset or cross-domain validation, limiting the ability to assess real-world generalizability.

4.6 Scalability and resource constraints

Scalability constraints were particularly evident in studies involving high-throughput networks, Internet of Things ecosystems, and edge computing environments. Several works noted that computationally intensive models introduce latency and resource overheads that hinder real-time detection. Memory consumption and energy efficiency were also identified as limiting factors, especially for resource-constrained devices. These constraints often necessitate trade-offs between detection accuracy and operational feasibility.

4.7 Privacy preservation limitations

Privacy preservation emerged as an underexplored area in the reviewed literature. Although some studies proposed privacy-aware approaches, such as federated learning or encrypted data processing, few provided comprehensive evaluations of privacy–utility trade-offs. In addition, many proposed techniques introduced additional communication and computational overheads, raising concerns about scalability. As a result, the practical adoption of privacy-preserving machine learning in cybersecurity remains limited.

5. Discussion

The results indicate that a significant portion of existing research on machine learning–based cybersecurity places disproportionate emphasis on improving algorithmic performance under controlled experimental conditions, while comparatively little attention is given to evaluating reliability in real-world operational environments. Many commonly used evaluation methodologies rely on static, curated datasets and simplified threat models that fail to reflect the complexity of real systems, where data is noisy, incomplete, and continuously evolving. In practice, cyber threats adapt rapidly, and attackers actively probe and manipulate detection mechanisms to evade defenses. As a result, the high detection accuracy and performance metrics reported in academic studies often overestimate the true effectiveness of these models once they are deployed in live networks.

This mismatch between experimental results and operational performance highlights several critical gaps in current research practices. Key factors such as resilience to adversarial manipulation, robustness under changing threat conditions, and the ability to operate consistently across diverse network environments are frequently underexplored. In addition, limited model interpretability remains a major obstacle to real-world adoption.

Security analysts must be able to understand, validate, and act upon system outputs, particularly in high-stakes environments such as security operations centers. When model decisions cannot be easily explained, trust is reduced, incident response is delayed, and the practical value of automated detection systems is diminished. Scalability is another concern that is often treated as secondary, despite being essential for deployment in large, heterogeneous, and resource-constrained infrastructures.

These findings suggest the need for a fundamental shift in how research on machine learning for cybersecurity is conducted and evaluated. Rather than prioritizing marginal improvements in benchmark accuracy, future work should focus on developing systems that maintain effectiveness in adversarial and dynamic settings. This includes designing models that can adapt to evolving attack strategies, incorporating realistic evaluation frameworks that simulate operational conditions, and explicitly measuring factors such as robustness, interpretability, computational efficiency, and generalization. Transparent and explainable models should be treated as essential components of security systems, enabling analysts to understand decision logic and respond effectively to detected threats.

Addressing such challenges requires collaboration that is interdisciplinary. Machine learning researchers, cybersecurity practitioners, and system engineers must work together from the early stages of system design to ensure that proposed solutions align with operational requirements and constraints. Such collaboration can help bridge the gap between theoretical advances and practical deployment by integrating domain knowledge, system-level considerations, and human factors into model development. Without this coordinated effort, progress in machine learning–based cybersecurity is likely to remain confined to academic environments, limiting its impact and preventing the widespread deployment of reliable, maintainable, and trustworthy security solutions in real-world systems.

Conclusion

Machine learning holds considerable promise for strengthening cyber threat detection; however, existing research has not yet enabled reliable and widespread deployment. Ongoing limitations related to adversarial vulnerability, model interpretability, data quality, generalization, scalability, and privacy continue to undermine real-world effectiveness. By reframing these challenges as interconnected barriers to deployment, this paper emphasizes the importance of aligning machine learning advancements with practical cybersecurity requirements. Addressing these issues is critical for developing robust, transparent, and deployable security systems capable of safeguarding modern digital infrastructure.

References

1. Al-Fawareh, M., Abu-Khalaf, J., Szewczyk, P., & Kang, J. J. (2023). Malware botnet detection using deep reinforcement learning in IoT networks. *IEEE Internet of Things Journal, 11*(1), 1–12. https://doi.org/10.1109/JIOT.2023.3324053
2. El Husseini, F., Noura, H. N., Salman, O., & Chehab, A. (2024). Advanced machine learning approaches for zero-day attack detection: A review. *IEEE Computer Security and Network Security Conference Proceedings*.
3. Yan, H., Li, X., Zhang, W., Wang, R., & Lin, X. (2024). Automatic evasion of machine learning-based network intrusion detection systems. *IEEE Transactions on Dependable and Secure Computing, 21*(3), 1–15.