

Artificial Intelligence in Cybersecurity: A Review of Machine Learning-Based Threat Detection Techniques, Challenges, and Future Directions

Dr. Amrish Dhawan¹, Karmjit Kaur², Amninder Kaur³

Director, Universal College, Patran, Punjab – 147105, India

HOD, Universal College, Patran, Punjab – 147105, India

Assistant Professor, Universal College, Patran, Punjab – 147105, India

1. Introduction

The increasing dependence on digital technologies has made cybersecurity an important concern for organisations, businesses, educational institutions, and government systems. The widespread adoption of cloud computing, connected devices, and online services has expanded the range of potential cyber threats, including malware, phishing, unauthorised access, and network intrusion. Traditional security mechanisms often rely on predefined rules and known attack signatures, which may limit their ability to identify unfamiliar or evolving threats.

Artificial intelligence (AI), particularly machine learning (ML), offers additional capabilities for analysing security-related data and identifying suspicious patterns. Machine learning algorithms can classify network traffic, recognise unusual behaviour, and assist in distinguishing legitimate activities from potentially malicious events. Different approaches, including supervised learning, unsupervised learning, and deep learning, have been explored for intrusion detection and related cybersecurity applications (Momand et al., 2023; Issa et al., 2024).

Despite these developments, the effectiveness of machine learning-based threat detection depends on several factors, including data quality, algorithm selection, evaluation methods, and the characteristics of the security environment. A model trained on a particular dataset may not perform equally well under different network conditions or when encountering previously unseen attacks. Similarly, high classification accuracy does not necessarily guarantee practical effectiveness if a system generates excessive false-positive alerts or requires substantial computational resources.

Another concern is the vulnerability of machine learning systems to adversarial manipulation. Attackers may attempt to modify inputs, influence training data, or exploit weaknesses in a model to evade detection.

Vassilev et al. (2024) provide a taxonomy of adversarial machine learning attacks and mitigation strategies, highlighting the importance of considering model security and robustness alongside predictive performance.

Against this background, this paper examines machine learning-based cybersecurity threat detection techniques, focusing on their principal approaches, potential applications, limitations, and practical challenges. The review also considers how evaluation practices and model robustness influence the reliability of these systems.

1.1 Research Objectives

The objectives of this study are:

1. To examine the principal machine learning techniques used for cybersecurity threat detection.
2. To discuss their applications in network intrusion detection and related security tasks.
3. To analyse the strengths and limitations of different detection approaches.
4. To identify challenges associated with false-positive alerts, generalisation, adversarial manipulation, and practical deployment.
5. To outline potential directions for improving the reliability and robustness of machine learning-based cybersecurity systems.

1.2 Research Questions

The study addresses the following research questions:

- **RQ1:** What machine learning techniques are used for cybersecurity threat detection?
- **RQ2:** What are the principal strengths and limitations of these techniques?
- **RQ3:** What challenges affect the reliability and practical deployment of machine learning-based detection systems?
- **RQ4:** What research directions could improve their robustness and ability to detect unfamiliar threats?

2. Review of Related Literature

Machine learning has become an important approach to cybersecurity threat detection, particularly in network intrusion detection and anomaly identification. Existing research examines different algorithms, datasets, evaluation methods, and security limitations. The following review summarises relevant literature and establishes the focus of the present study.

2.1 Machine Learning Approaches to Intrusion Detection

Momand et al. (2023) reviewed developments in machine learning-based intrusion detection, with particular attention to deep learning techniques, datasets, and attack taxonomies. Their work provides an overview of the approaches used to identify suspicious network activities and highlights the importance

of dataset characteristics in evaluating detection performance. The availability of representative data and the coverage of different attack categories are important considerations when assessing these techniques.

2.2 Evaluation and Methodological Limitations

Issa et al. (2024) examined research trends, algorithms, datasets, evaluation methods, and limitations in intrusion detection systems. Their review highlights the importance of considering the conditions under which detection techniques are evaluated. Differences in datasets, preprocessing methods, and evaluation metrics can make direct comparisons between algorithms difficult. Consequently, practical assessment should consider measures such as precision, recall, and false-positive rate alongside accuracy.

2.3 Adversarial Robustness in Cybersecurity

Vassilev et al. (2024) presented a taxonomy of adversarial machine learning attacks and mitigation strategies. Their work explains how deliberate manipulation of model inputs or training processes can affect machine learning system behaviour. In cybersecurity, such manipulation may enable attackers to evade detection or influence model predictions. These risks highlight the importance of considering adversarial robustness alongside conventional detection performance.

2.4 Research Gap and Focus of the Present Study

The reviewed literature addresses intrusion detection techniques, evaluation limitations, dataset characteristics, and adversarial vulnerabilities from complementary perspectives. Building on these contributions, the present study brings these themes together to examine machine learning-based threat detection in terms of detection capability, generalisation to unfamiliar attacks, false-positive alerts, and adversarial robustness. It also considers computational requirements and practical deployment challenges to identify potential directions for developing more reliable cybersecurity detection systems.

3. Research Methodology

This study adopts a qualitative, literature-based review approach to examine artificial intelligence, particularly machine learning, techniques used in cybersecurity threat detection. The review draws on published scholarly literature and technical reports addressing intrusion detection systems, machine learning algorithms, evaluation practices, and adversarial machine learning.

The selected literature is examined thematically, with particular attention to four major approaches: supervised learning, unsupervised learning, deep learning, and hybrid or ensemble methods. The analysis considers their detection capabilities, data requirements, potential limitations, and suitability for different cybersecurity applications.

The review also examines key factors affecting the reliability of machine learning-based threat detection, including dataset quality, generalisation to unfamiliar attacks, false-positive alerts, computational

requirements, interpretability, and adversarial vulnerabilities. These themes are used to compare the approaches discussed in the literature and identify potential directions for further research.

The study does not present new experimental results or claim to provide an exhaustive assessment of all published research. Instead, it offers a focused qualitative synthesis of the selected literature to develop an understanding of current detection techniques, their practical challenges, and considerations for improving the reliability and robustness of cybersecurity systems.

4. Machine Learning-Based Threat Detection Techniques

Machine learning-based threat detection uses network traffic, system events, and other security-related data to identify potentially malicious activities. The principal approaches include supervised learning, unsupervised learning, deep learning, and hybrid or ensemble methods. Each approach differs in its data requirements, detection capabilities, and computational complexity (Momand et al., 2023; Issa et al., 2024).

4.1 Supervised Learning

Supervised learning uses labelled training data to classify observations into predefined categories, such as normal and malicious network traffic. Common algorithms used in intrusion detection include Decision Trees, Support Vector Machines (SVM), Logistic Regression, and Naïve Bayes (Momand et al., 2023). These algorithms can support the classification of known threat patterns. However, their effectiveness depends on the representativeness of the training data, and they may struggle to identify attacks that differ substantially from previously observed examples.

4.2 Unsupervised Learning

Unsupervised learning identifies patterns and unusual behaviour without requiring predefined labels for every observation. Clustering algorithms, such as K-means, and anomaly detection methods can help identify network activities that deviate from established patterns (Momand et al., 2023).

This approach is useful when labelled attack data is limited. However, unusual behaviour does not necessarily indicate malicious activity, and legitimate changes in network usage may also trigger alerts.

4.3 Deep Learning

Deep learning uses multilayer neural networks to learn complex patterns from data. Techniques discussed in intrusion detection research include Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), recurrent neural networks (RNNs), and autoencoders (Momand et al., 2023). Issa et al. (2024) also identify CNNs among the approaches used in intrusion detection research.

These architectures support different analytical tasks. CNNs can learn features for classification, RNNs can model sequential information, and autoencoders can learn data representations useful for anomaly detection. However, deep learning models may require substantial computational resources and can be difficult to interpret.

4.4 Hybrid and Ensemble Approaches

Hybrid approaches combine different detection methods or integrate machine learning with conventional security mechanisms. For example, a system may combine rule-based detection with anomaly detection to identify recognised attack patterns as well as unusual network behaviour.

Ensemble learning combines predictions from multiple models to exploit complementary strengths. Although these approaches may improve detection capabilities in suitable settings, they can also increase computational demands and system complexity. Their effectiveness must therefore be established through appropriate evaluation rather than assumed.

Overall, the choice of a machine learning approach depends on the detection objective, available data, and operational requirements. No single technique is universally suitable for every cybersecurity environment, making comparative evaluation important when selecting a detection strategy (Momand et al., 2023; Issa et al., 2024).

5. Comparative Analysis of Detection Techniques

Machine learning-based threat detection techniques differ in their data requirements, detection capabilities, and operational limitations. Comparing these characteristics helps clarify the suitability of different approaches for cybersecurity applications (Momand et al., 2023; Issa et al., 2024).

5.1 Comparison of Major Approaches

Technique	Data requirements	Potential strengths	Key limitations
Supervised learning	Labelled training data	Classifies predefined threat categories	May struggle with unfamiliar or underrepresented attacks
Unsupervised learning	Primarily unlabelled data	Identifies unusual patterns without extensive attack labels	May generate false-positive alerts
Deep learning	Suitable data and computational resources	Learns complex patterns and representations	Can be computationally expensive and difficult to interpret
Hybrid and ensemble methods	Depends on the combined techniques	Integrates complementary detection capabilities	May increase system complexity and maintenance requirements

The comparison indicates that each approach involves trade-offs. The most suitable technique depends on the detection objective, data availability, and operational environment. Combining methods may offer complementary capabilities, but improved performance must be demonstrated through evaluation.

5.2 Evaluation of Detection Performance

Accuracy alone may provide an incomplete assessment, particularly when datasets contain substantially more benign observations than malicious ones. Additional metrics help explain how a detection system performs under different conditions.

Precision measures the proportion of flagged observations that are genuinely malicious, while recall measures the proportion of actual malicious observations correctly identified. The F1-score balances precision and recall, and the false-positive rate measures how frequently legitimate observations are incorrectly classified as malicious.

The importance of each metric depends on the application. Excessive false-positive alerts can increase the workload of security analysts, whereas missed attacks may leave systems vulnerable. Evaluation should therefore consider both detection effectiveness and the consequences of classification errors.

5.3 Generalisation and Practical Reliability

A detection model may perform differently when applied to network environments or attack patterns that differ from its training data. Evaluation across different datasets, time periods, or network conditions can provide additional evidence about its ability to generalise.

Adversarial robustness is another consideration because deliberate manipulation may affect model predictions even when conventional test performance is strong (Vassilev et al., 2024). A practical assessment should therefore consider detection metrics, generalisation, computational requirements, and resilience to adversarial manipulation rather than relying on accuracy alone.

6. Challenges in Machine Learning-Based Cybersecurity

Despite their potential, machine learning-based detection systems face challenges that can affect their reliability and practical deployment. These include unfamiliar attack patterns, limitations in training data, false-positive alerts, adversarial vulnerabilities, and computational constraints.

6.1 Detection of Unfamiliar Attacks

Models trained on previously observed attacks may fail to recognise threats that differ substantially from their training examples. Changes in attack techniques and network behaviour can reduce detection effectiveness, highlighting the importance of evaluating models beyond familiar data patterns (Momand et al., 2023; Issa et al., 2024).

6.2 Data Quality and Representation

Incomplete labels, imbalanced attack categories, and unrepresentative datasets can influence model training and evaluation. A model developed using limited or outdated data may not perform consistently

in a different operational environment. Representative data and careful evaluation are therefore important for assessing detection reliability.

6.3 False-Positive Alerts

False positives occur when legitimate activities are incorrectly classified as malicious. A high volume of unnecessary alerts can burden security analysts and make genuine threats more difficult to identify. Detection systems must therefore balance the identification of malicious activity with the need to limit incorrect alerts.

6.4 Adversarial Vulnerabilities

Machine learning-based threat detection systems may be targeted through deliberate manipulation of inputs or training data. For example, in an **evasion attack**, an attacker may modify network traffic patterns so that malicious activity resembles legitimate behaviour, potentially avoiding detection. In a **data poisoning attack**, manipulated or incorrectly labelled samples may be introduced into training data, potentially influencing the model to misclassify malicious activity.

Vassilev et al. (2024) describe adversarial machine learning attacks and mitigation strategies, highlighting the need to address vulnerabilities throughout the machine learning lifecycle. These examples illustrate why conventional performance metrics alone may not establish a model's resilience to deliberate manipulation. Appropriate adversarial testing, data validation, and mitigation measures should therefore complement standard detection evaluation.

6.5 Computational and Interpretability Constraints

Complex models, particularly some deep learning approaches, may require substantial processing power and memory, limiting their suitability for resource-constrained environments. In addition, models whose predictions are difficult to interpret may complicate security investigations. Balancing detection capability with computational efficiency and interpretability remains an important practical consideration.

7. Future Research Directions

The challenges discussed in this review suggest several directions for improving the reliability and practical applicability of machine learning-based cybersecurity threat detection. Future research should focus on generalisation, data quality, detection accuracy, adversarial robustness, and operational efficiency.

7.1 Improving Detection of Unfamiliar Attacks

Future studies should investigate methods that can identify attacks not adequately represented in training data. Approaches such as anomaly detection and continual learning warrant further evaluation under changing network conditions to determine their effectiveness in recognising emerging threats.

7.2 Developing Representative Evaluation Datasets

Research should examine how dataset diversity, label quality, and changing attack distributions influence detection performance. Evaluating models across multiple datasets and time periods could provide a clearer understanding of their generalisation capabilities and practical limitations.

7.3 Reducing False-Positive Alerts

Future work should explore methods that reduce false-positive alerts while maintaining effective detection of malicious activity. Context-aware classification, feature selection, and appropriate decision thresholds are potential areas for investigation. Their effectiveness should be assessed using relevant detection metrics and operational requirements.

7.4 Strengthening Adversarial Robustness

Further research should investigate the resilience of detection models to deliberate manipulation. Potential approaches include adversarial training, improved training-data validation, and systematic robustness testing. Building on the framework presented by Vassilev et al. (2024), these methods should be evaluated against clearly defined attack scenarios rather than relying solely on conventional classification performance.

7.5 Improving Efficiency and Explainability

Future studies should explore detection models that balance predictive performance with computational efficiency and interpretability. Lightweight architectures may support deployment in resource-constrained environments, while explainable methods could help security analysts understand model predictions and investigate suspicious activity.

Overall, future research should prioritise realistic evaluation, measurable improvements in robustness, and suitability for operational environments. These efforts could help establish the conditions under which different machine learning techniques provide reliable support for cybersecurity threat detection.

8. Conclusion

Artificial intelligence, particularly machine learning, offers several approaches to cybersecurity threat detection, including supervised learning, unsupervised learning, deep learning, and hybrid methods. The suitability of these techniques depends on the detection task, available data, and operational requirements. This review highlights key challenges affecting their practical reliability, including unfamiliar attacks, data limitations, false-positive alerts, adversarial vulnerabilities, and computational constraints. These considerations indicate that detection systems should be assessed not only through conventional performance metrics but also through their generalisation, robustness, and suitability for real-world deployment.

Future research should prioritise realistic evaluation, improved detection of emerging threats, stronger adversarial resilience, and a balance between performance and operational efficiency. Such efforts could support the development of more dependable machine learning-based cybersecurity systems.

REFERENCES:

1. Issa, M. M., Aljanabi, M., & Muhialdeen, H. M. (2024). Systematic literature review on intrusion detection systems: Research trends, algorithms, methods, datasets, and limitations. *Journal of Intelligent Systems*, 33(1), Article 20230248. <https://doi.org/10.1515/jisys-2023-0248>
2. Momand, A., Jan, S. U., & Ramzan, N. (2023). A systematic and comprehensive survey of recent advances in intrusion detection systems using machine learning: Deep learning, datasets, and attack taxonomy. *Journal of Sensors*, 2023, Article 6048087. <https://doi.org/10.1155/2023/6048087>
3. Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (NIST AI 100-2e2023). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-2e2023>