IJVAST | Vol. 2, Issue 2 | June 2026 | Open Access — Peer Reviewed Review Article

A Review Paper on AI-Based Cyber Attack Detection System

Kanchan Minde[1], Rutuja Bhangare[2]

Department of Computer Science and Engineering

Agasti Institute of Management, Akole, Maharashtra, India

Abstract —*The growing scale, sophistication, and frequency of cyber attacks have exposed the limitations of traditional signature-based intrusion detection systems, which struggle to identify novel or evolving attack patterns not present in existing signature databases. Artificial Intelligence (AI) and Machine Learning (ML) based approaches offer a complementary detection paradigm capable of learning patterns of normal and malicious behavior directly from network traffic and system data, enabling detection of previously unseen attacks. This review paper examines AI-based approaches to cyber attack detection, covering classical machine learning techniques, deep learning architectures including convolutional and recurrent neural networks, and anomaly-based detection methods that flag deviations from learned normal-behavior baselines. The paper reviews the range of attack types addressed by these systems — including denial-of-service, network intrusion, malware-related traffic, and insider threats — and examines commonly used benchmark datasets, feature engineering practices, and detection evaluation metrics. A comparative analysis of representative studies highlights reported detection accuracy and persistent limitations, including class imbalance, susceptibility to zero-day attacks, adversarial evasion, and the operational challenge of maintaining low false-positive rates in real-time deployment. The review concludes with future research directions spanning federated and privacy-preserving detection, explainable AI for security operations, adversarial robustness, and lightweight models for edge and IoT deployment.*

Index Terms—*Cyber Attack Detection, Machine Learning, Deep Learning, Intrusion Detection System, Anomaly Detection, Network Security, Adversarial Machine Learning, Explainable AI.*

I. Introduction

Cyber attacks targeting networks, applications, and connected devices have grown substantially in scale and sophistication, ranging from denial-of-service floods and network intrusions to malware-driven compromise and insider misuse of legitimate access. Traditional intrusion detection systems rely primarily on signature-based matching, comparing observed traffic or system behavior against a database of known attack patterns; while effective against previously catalogued threats, this approach is inherently limited in detecting novel or evolving attacks that lack a matching signature [1],[4].

Artificial Intelligence and Machine Learning based detection approaches address this limitation by learning statistical or behavioral patterns of normal and malicious activity directly from data, enabling generalization to attack variants not explicitly present in training data. Early applications of data mining and machine learning to intrusion detection demonstrated the feasibility of this approach, though also revealed practical challenges around false-positive rates and the difficulty of validating detection systems under realistic, non-stationary network conditions [3],[4]. Deep learning architectures have since been applied to further improve detection accuracy by automatically learning hierarchical feature representations from raw or lightly processed network traffic data [5],[8].

This review examines AI-based approaches to cyber attack detection published primarily between 2009 and 2022. The primary contributions are: (1) a structured review of machine learning and deep learning approaches to attack detection; (2) an analysis of the categories of cyber attacks addressed

III. AI/ML Approaches for Cyber Attack Detection

A. Classical Machine Learning Techniques

Early AI-based intrusion detection research applied classifiers such as Decision Trees, Support Vector Machines, Naive Bayes, and Random Forests to hand-engineered network traffic features, achieving reasonable detection accuracy on benchmark datasets while requiring domain expertise for effective feature selection [3].

B. Deep Learning Architectures

Deep neural networks, including feed-forward architectures and convolutional neural networks, have been applied to learn discriminative features directly from network traffic data, generally improving detection accuracy over classical ML approaches on standard intrusion detection benchmarks [5],[6].

C. Recurrent and Sequence-Based Models

Given the inherently sequential nature of network traffic, recurrent architectures such as Long Short-Term Memory (LSTM) networks have been applied to capture temporal dependencies in traffic flows, improving detection of attacks that unfold over multiple packets or sessions [7],[8].

D. Anomaly Detection Methods

Unsupervised and semi-supervised anomaly detection techniques, including clustering and autoencoder-based reconstruction error methods, are used to flag deviations from learned normal-behavior baselines without requiring labeled attack data, an important capability given limited availability of labeled attack samples for novel threats

by AI-based detection systems; (3) a review of benchmark datasets and feature engineering practices; (4) a comparative analysis of representative studies; and (5) identification of open challenges and future research directions.

II. Background

A. Signature-Based vs. Anomaly-Based Detection

Signature-based systems detect attacks by matching observed activity against known attack signatures, offering low false-positive rates for known threats but no protection against novel attacks, while anomaly-based systems model normal behavior and flag deviations, offering broader coverage at the cost of higher false-positive rates [14].

B. The Case for AI-Based Detection

AI-based detection systems are positioned as a complement to signature-based methods, learning from historical attack and benign traffic data to generalize detection capability to attack variants and, to a limited extent, previously unseen attack types [3].

IV. Comparison of Approaches

Table I summarizes classical ML, deep learning, and anomaly-based approaches to AI-based cyber attack detection. Classical ML approaches remain the most interpretable and computationally lightweight but generally show lower detection accuracy on complex, high-dimensional traffic data relative to deep learning approaches, which in turn require larger training datasets and offer reduced interpretability.

TABLE I — Comparison of AI/ML Approaches for Cyber Attack Detection

Approach	Key Technique	Interpretability	Data Requirement	Novel-Attack Generalization
Classical ML (SVM/RF/DT) [3]	Hand-crafted traffic features	High	Low-Moderate	Low-Moderate
Deep Neural Networks [5]	End-to-end feature learning	Low	High	Moderate
Recurrent/LSTM Models [7],[8]	Sequential traffic modeling	Low	High	Moderate-High
Anomaly Detection [14]	Deviation from normal-behavior baseline	Moderate	Moderate	High (for behavioral novelty)
Hybrid/Ensemble	Multi-model aggregation	Low-Moderate	High	High

V. Categories of Cyber Attacks Addressed

A. Denial-of-Service and Distributed Denial-of-Service

AI-based detection systems are widely applied to identify DoS/DDoS traffic patterns, characterized by abnormal traffic volume or connection request rates, using both classical ML classifiers and deep learning models trained on flow-level traffic features [3],[6].

B. Network Intrusion and Probing Attacks

Detection of unauthorized access attempts and network reconnaissance/probing activity remains one of the most extensively studied applications of AI-based detection, forming the core task addressed by widely used benchmark datasets [1],[2].

C. Malware-Related Network Traffic

AI-based systems can flag network traffic patterns associated with malware command-and-control communication or data exfiltration, complementing endpoint-based malware detection with a network-traffic-centric detection perspective.

D. Insider Threats and Behavioral Anomalies

Detecting insider threats requires modeling legitimate user behavior baselines and flagging deviations, such as unusual data access patterns, a task well suited to anomaly-based rather than signature-based detection approaches given the absence of a fixed attack signature [14].

VI. Comparative Analysis

Table II compares representative studies on AI-based cyber attack detection across dataset, technique, and reported performance. Deep learning approaches generally report higher detection accuracy on benchmark datasets relative to classical ML baselines, a pattern [14].

VII. Datasets and Feature Engineering

A. Benchmark Intrusion Detection Datasets

Widely used benchmark datasets include NSL-KDD, a refined version of the original KDD Cup 99 dataset addressing known redundancy issues [1], CICIDS2017, which provides more contemporary traffic patterns and attack types collected under controlled network conditions [2], and UNSW-NB15, which combines real modern normal traffic with synthetically generated contemporary attack behaviors [13].

B. Feature Selection and Engineering

Effective detection depends on selecting informative traffic features — including flow duration, packet size statistics, and protocol-specific attributes — with feature selection techniques used to reduce dimensionality and improve both accuracy and computational efficiency [3].

C. Raw Traffic vs. Flow-Level Analysis

Detection systems vary in whether they operate on flow-level aggregated statistics or attempt to learn directly from raw packet sequences, with the latter approach, often paired with deep learning, offering potentially richer feature representations at higher computational cost [5].

D. Explainability in Security Contexts

Given the operational consequences of detection decisions, explainability techniques that surface which features drove a given detection are increasingly emphasized to support security analyst trust and enable actionable incident response.

VIII. Benchmark Dataset Characteristics

Table III summarizes key characteristics of widely used intrusion

confirmed across multiple datasets in comparative evaluation studies [11], though critical analyses caution that benchmark performance does not always translate to comparable performance under real-world, non-stationary network conditions [4].

detection benchmark datasets, providing context for the comparative analysis of representative studies in Section VI.

TABLE II — Comparative Analysis of Representative Studies

Author & Year	Dataset	Technique	Task	Reported Performance
Javaid et al. [5] (2016)	NSL-KDD	Deep learning (self-taught)	Attack classification	Improved accuracy vs. classical ML baselines
Yin et al. [8] (2017)	NSL-KDD	Recurrent neural network	Binary/multi-class classification	High accuracy on benchmark test set
Kim et al. [7] (2016)	KDD-based traffic	LSTM classifier	Intrusion classification	Effective modeling of sequential traffic patterns
Vinayakumar et al. [6] (2019)	Multiple benchmark datasets	Deep learning survey/evaluation	Multi-dataset evaluation	Consistent gains from deep architectures across datasets
Sommer & Paxson [4] (2010)	General network traffic	Critical analysis	Methodological critique	Highlighted gap between benchmark and real-world performance
Apruzzese et al. [10] (2018)	General ML-IDS studies	Critical/empirical analysis	Effectiveness assessment	Identified operational limitations of ML-based detection

Dataset	Year	Attack Categories	Notes
KDD Cup 99 / NSL-KDD [1]	1999 / 2009 (refined)	DoS, Probe, R2L, U2R	NSL-KDD removes redundant records present in original KDD Cup 99
CICIDS2017 [2]	2017	DoS, DDoS, Brute Force, Infiltration, Botnet, Web Attacks	Contemporary traffic patterns, labeled flow features
UNSW-NB15 [13]	2015	Fuzzers, Backdoors, DoS, Exploits, Generic, Reconnaissance, and others	Blend of real normal traffic and synthetic contemporary attacks
Various benchmarks (survey)	2019 (survey)	Multiple	Comprehensive survey of network-based IDS datasets and limitations [12]

TABLE III — Benchmark Intrusion Detection Datasets

IX. Key Challenges

A. Class Imbalance

Benchmark and real-world network traffic datasets are heavily skewed toward benign traffic, with attack samples, particularly for rare attack types, substantially underrepresented, causing models to underperform on minority attack classes unless mitigated through resampling or class-weighted training [3].

B. Zero-Day Attack Detection

While AI-based systems generalize better than signature-based methods, detecting genuinely novel, previously unseen attack techniques remains fundamentally challenging, since models are ultimately trained on historical attack patterns [4].

C. Adversarial Machine Learning

AI-based detection systems are themselves susceptible to adversarial manipulation, where attackers craft traffic specifically designed to evade detection by exploiting model decision boundaries, a vulnerability well documented in the broader adversarial machine learning literature [9].

D. Real-Time Detection Requirements

Operational deployment requires detection latency low enough to support real-time or near-real-time response, constraining model complexity and motivating research into lightweight architectures

X. Future Research Directions

A. Federated and Privacy-Preserving Detection

Federated learning approaches that train shared detection models across organizations without centralizing raw network traffic data offer a promising direction for improving detection generalization while respecting data privacy and confidentiality constraints.

B. Explainable AI for Security Operations

Continued development of explainability techniques tailored to security contexts is needed to help analysts interpret and act on AI-generated detection alerts, particularly for deep learning models whose decision processes are otherwise opaque.

C. Adversarial Robustness

Research into training techniques that improve model robustness against adversarial evasion, such as adversarial training and robust feature representations, is an important direction given attacker incentive to specifically target AI-based detection systems [9].

D. Lightweight Models for Edge and IoT Deployment

Developing lightweight, low-latency detection models suitable for deployment on resource-constrained edge devices and IoT gateways is increasingly important given the growing scale of connected device deployment and the corresponding expansion of

suitable for high-throughput network environments.

E. False-Positive Rate Management

High false-positive rates impose significant operational burden on security analysts and can lead to alert fatigue, underscoring the importance of evaluating detection systems on precision and false-positive rate alongside overall accuracy [4].

the network attack surface.

XI. Conclusion

This review has examined AI-based approaches to cyber attack detection, spanning classical machine learning, deep learning, and anomaly-based methods, along with the datasets, feature engineering practices, and evaluation considerations underlying this research area. The evidence indicates that AI-based detection systems offer meaningful improvements over purely signature-based approaches in generalizing to attack variants, though persistent challenges including class imbalance, zero-day attack detection, adversarial evasion, and false-positive rate management remain significant barriers to reliable real-world deployment. Future research in federated learning, explainable AI, adversarial robustness, and lightweight edge-deployable models is expected to be central to advancing AI-based cyber attack detection toward robust, operationally trustworthy security tools.

XII. References

[1] Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A.A. (2009). A detailed analysis of the KDD CUP 99 data set. Proc. IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 1-6.

[2] Sharafaldin, I., Lashkari, A.H., and Ghorbani, A.A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proc. International Conference on Information Systems Security and Privacy (ICISSP), 108-116.

[3] Buczak, A.L., and Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153-1176.

[4] Sommer, R., and Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. Proc. IEEE Symposium on Security and Privacy, 305-316.

[5] Javaid, A., Niyaz, Q., Sun, W., and Alam, M. (2016). A deep learning approach for network intrusion detection system. Proc. 9th EAI International Conference on Bio-inspired Information and Communications Technologies (BICT), 21-26.

[6] Vinayakumar, R., Alazab, M., Soman, K.P., Poornachandran, P., Al-Nemrat, A., and Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. IEEE Access, 7, 41525-41550.

[7] Kim, J., Kim, J., Thu, H.L.T., and Kim, H. (2016). Long short term memory recurrent neural network classifier for intrusion detection. Proc. International Conference on Platform Technology and Service (PlatCon), 1-5.

[8] Yin, C., Zhu, Y., Fei, J., and He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access, 5, 21954-21961.

[9] Goodfellow, I.J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. Proc. International Conference on Learning Representations (ICLR).

[10] Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., and Marchetti, M. (2018). On the effectiveness of machine learning and deep learning for cyber security. Proc. 10th International Conference on Cyber Conflict (CyCon), 371-390.

[11] Ferrag, M.A., Maglaras, L., Moschoyiannis, S., and Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications, 50, 102419.

[12] Ring, M., Wunderlich, S., Scheuring, D., Landes, D., and Hotho, A. (2019). A survey of network-based intrusion detection data sets. Computers & Security, 86, 147-167.

[13] Moustafa, N., and Slay, J. (2015). UNSW-NB15: a comprehensive data set for network intrusion detection systems. Proc. Military Communications and Information Systems Conference (MilCIS), 1-6.

[14] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1-58.

ISSN: Applied For (2026) | DOI: 10.XXXXX/IJVAST.2026.XXXXX | Open Access — CC BY 4.0 License